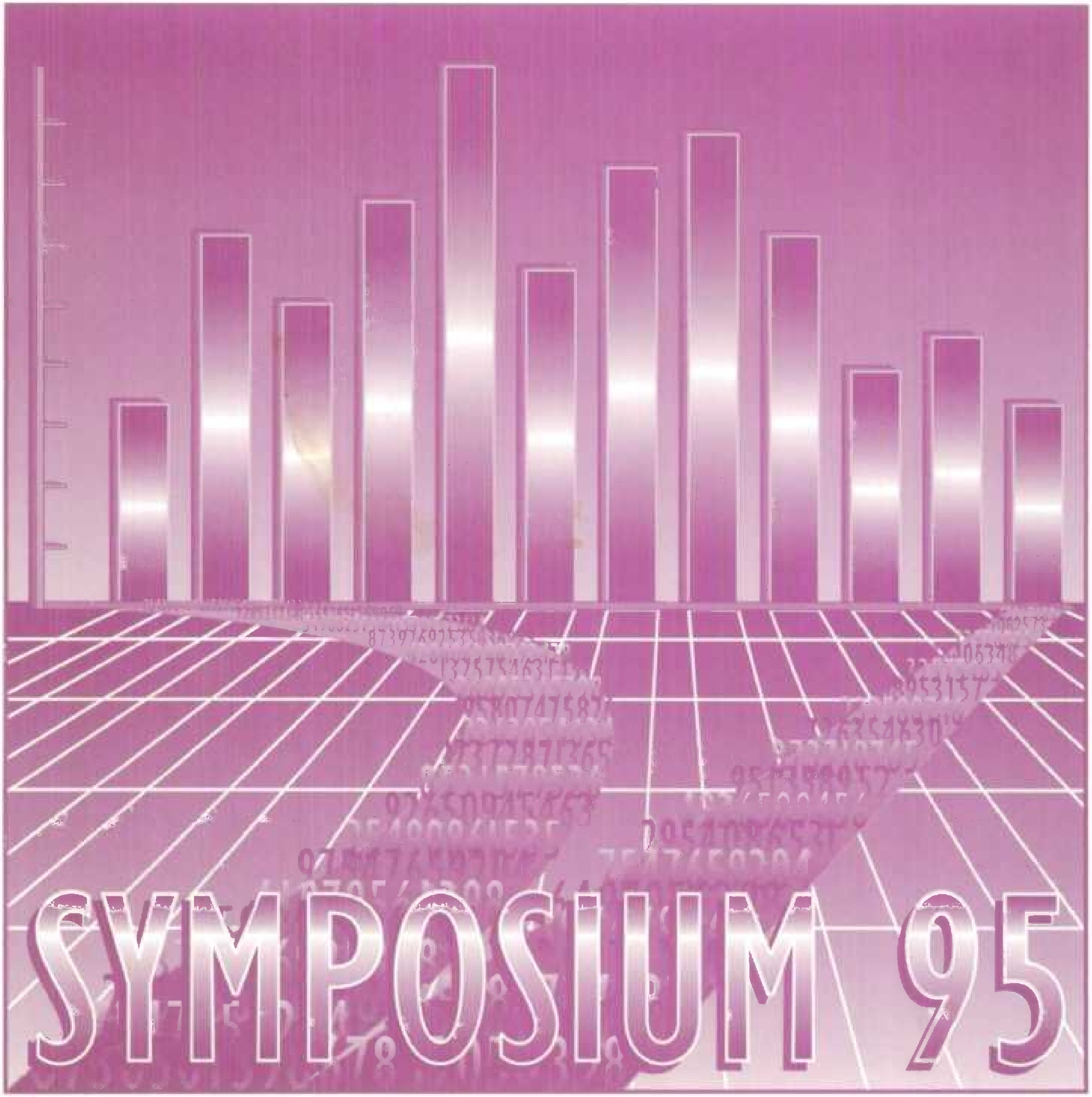
c.3

Catalogue no. 11-522-XPE

SYMPOSIUM 95

From Data to Information Methods and Systems

PROCEEDINGS



Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this publication and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone (613) 951-8615) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http//www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialling area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

How to order publications

Statistics Canada publications may be purchased from local authorized agents and other community bookstores, the Statistics Canada Regional Reference Centres, or from:

Statistics Canada
Operations and Integration Division
Circulation Management
120 Parkdale Avenue
Ottawa, Ontario
K1A 0T6

Telephone: (613) 951-7277
Fax: (613) 951-1584
Toronto (credit card only): (416) 973-8018
Internet: order@statcan.ca

Standards of service to the public

To maintain quality service to the public, Statistics Canada follows established standards covering statistical products and services, delivery of statistical information, cost-recovered services and services to respondents. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Methodology Branch

SYMPOSIUM 95

From Data to Information Methods and Systems

PROCEEDINGS

Published by authority of the Minister responsible for Statistics Canada

October 1996

Price: Canada: $53.00
United States: US$53.00
Other Countries: US$53.00

Catalogue no. 11-522-XPE

ISBN 0-660-16585-6

Ottawa

La version française de cette publication est disponible sur demande (n° 11-522-XPF au catalogue).

Note of Appreciation

Canada owes the success of its statistical system to a long-standing cooperation involving Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Symposium '95, From Data to Information
(1995 : Ottawa, Ont.)
Symposium '95, From Data to Information :
methods and systems : proceedings.

Issued also in French under title: Symposium
'95, Des Données à l'Information : méthodes
et systèmes : recueil.
ISBN 0-660-16585-6
CS11-522-XPE

1. Statistical services -- Congresses. 2. Statistics --
Congresses. I. Statistics Canada. Methodology
Branch. II. Title.

H21 S95 1996 001.4'22
C96-98801-7

PREFACE

Symposium 95 was the twelfth in the series of international symposia on methodological issues sponsored by Statistics Canada. Each year the symposium focuses on a particular theme. This year, the emphasis was on the analysis and dissemination stages in the information development processes.

The 1995 symposium attracted close to 300 persons who met over three days in the Simon Goldberg Conference Centre in Ottawa. Presentations by academic and government statisticians, specialists in information processing and information management, data vendors and end users were heard. A total of 29 papers were presented by the invited speakers and panellists. Aside from translation and formatting, the papers submitted by the authors have been reproduced in these proceedings. Presentations by two of the panellists were transcribed from recordings and underwent minor editorial changes.

The organizers of Symposium 95 would like to acknowledge the contributions of many persons involved in the preparation of this volume and those who assisted them during the symposium in November. The committee would especially like to thank Josée Morel, Sophie Arsenault, Christine Larabie and Nick Budko for the many hours of preparing material and making arrangements for Symposium 95.

Naturally, the presenters at Symposium 95 deserve thanks for taking the time to put their ideas into written form. Publication of these proceedings also involved the efforts of many others. Processing of the manuscript was expertly handled by Christine Larabie with the assistance of Judy Clarke, Sandy Diloreto and Suzanne Fleury-Bertrand. Proofreading was done by a number of Statistics Canada methodologists: Jean-Luc Bernier, Alana Boltwood, René Boyer, Guylaine Dubreuil, Sylvie Gauthier, John Higginson, Tony LaBillois, Éric Langlet, Éric Lesage, Mary March, Josée Morel, Carole Morin, Sylvain Perron, Craig Seko, Michelle Simard, Jack Singleton and Larry Swain. Production of these proceedings was coordinated by Christine Larabie.

Statistics Canada's thirteenth annual symposium, to be held November 13 to 15, 1996 in Ottawa, will be preceded by a one-day workshop. Their theme will be nonsampling errors.

Symposium 95 Organizing Committee

John Berigan
Georgia Roberts
Jean Dumais
Jean-Louis Tambay

STATISTICS CANADA SYMPOSIUM SERIES

1984 -	Analysis of Survey Data
1985 -	Small Area Statistics
1986 -	Missing Data in Surveys
1987 -	Statistical Uses of Administrative Data
1988 -	The Impact of High Technology on Survey Taking
1989 -	Analysis of Data in Time
1990 -	Measurement and Improvement of Data Quality
1991 -	Spatial Issues in Statistics
1992 -	Design and Analysis of Longitudinal Surveys
1993 -	International Conference on Establishment Surveys
1994 -	Re-engineering for Statistical Agencies
1995 -	From Data to Information - Methods and Systems

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

Use the order form on this page to order additional copies of the proceedings of Symposium 95: From Data to Information - Methods and Systems. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 95 PROCEEDINGS
STATISTICS CANADA
BUSINESS SURVEY METHODS DIVISION
R.H. COATS BUILDING, 11th FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada" - Indicate on cheque or money order: Symposium 95 - Proceedings).

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH		_____ @ $10
1987 -	Les utilisations statistiques des données administratives - FRENCH		_____ @ $10
1987 -	SET OF 1 ENGLISH AND 1 FRENCH		_____ @ $12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL		_____ @ $10
1989 -	Analysis of Data in Time - BILINGUAL		_____ @ $15
1990 -	Measurement and Improvement of Data Quality - ENGLISH		_____ @ $18
1990 -	Mesure et amélioration de la qualité des données - FRENCH		_____ @ $18
1991 -	Spatial Issues in Statistics - ENGLISH		_____ @ $20
1991 -	Questions spatiales liées aux statistiques - FRENCH		_____ @ $20
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH		_____ @ $22
1992 -	Conception et analyse des enquêtes longitudinales - FRENCH		_____ @ $22
1993 -	International Conference on Establishment Surveys - ENGLISH (available in English only, published in U.S.A.)		_____ @ $58
1994 -	Re-engineering for Statistical Agencies - ENGLISH		_____ @ $53
1994 -	Restructuration pour les organismes de statistique - FRENCH		_____ @ $53
1995 -	From Data to Information - Methods and Systems - ENGLISH		_____ @ $53
1995 -	Des données à l'information - Méthodes et systèmes - FRENCH		_____ @ $53
	PLEASE ADD THE GOODS AND SERVICES TAX (7%) (Residents of Canada only)	$	_____
	TOTAL AMOUNT OF ORDER	$	_____

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME ____

ADDRESS ____

____

CITY ____________ PROV/STATE ____________ COUNTRY ____________

POSTAL CODE ________ TELEPHONE ____________ FAX ____________

For more information please contact John Kovar, Telephone (613) 951-8615, Facsimile (613) 951-1462, E-mail kovar@statcan.ca

FROM DATA TO INFORMATION - METHODS AND SYSTEMS

TABLE OF CONTENTS[1]

[1] In cases of joint authorship, the name of the presenter is shown **boldface**.

SESSION 8: Electronic Information Dissemination
Chairperson: M. Podehl, Statistics Canada

SESSION 9: Panel

OPENING REMARKS

OPENING REMARKS

G.J. Brackstone[1]

On behalf of Statistics Canada, welcome to this the 12th Symposium in a series that goes back to 1984. For those of you from out of town, welcome to Ottawa, and for those of you from outside the country, welcome to Canada.

Since its inception in 1984 these Symposia series have covered a variety of topics associated with survey-taking in the broadest sense. For example, we have dealt with Small Area Statistics, Statistical uses of Administrative Data, The Impact of Technology on Survey Taking, Design and Analysis of Longitudinal Surveys, and Establishment Surveys. Last year we tackled a broader topic that went beyond survey methodology and analysis when we addressed Re-engineering in Statistical Agencies. This year's topic is also a broad one. It takes us back to our first Symposium topic: Analysis of Survey Data, and broadens it to address the challenge of converting data held by statistical agencies into information valuable to users.

This year we are shifting the focus specifically to what one might call, without any derogatory intent, the rear end of the survey process - towards those activities that need to take place after the intensive operations of collection and processing are over and we have data which we want to ensure are used effectively.

In recent years, in many statistical agencies, the attention being paid to these output-oriented activities has increased considerably. Two main factors have driven this increased attention I would suggest. The first was the recognition that statistical agencies needed to strengthen their client orientation if they were to retain their relevance and therefore their support and funding; and the second was the financial pressure that dictated that data already collected had to be fully used and costs associated with dissemination activities fully recovered.

Since the 1960s, the notion of each single survey, or each single data collection vehicle, leading to a single, or maybe several, publications as the output of that survey has been supplanted. The first progression was to the idea that the basic output of a survey was not a publication but a database from which a publication might be produced, but from which also might be produced supplementary tables or analyses on demand. In other words a database and a retrieval capability were the principal outputs of a survey. That view too is now itself being supplanted by the notion that a survey should be thought of as contributing to a corporate base of information, which may contain data from many different sources, and from which information can be retrieved in a common integrated way - a corporate database that provides the foundation for an information service utilizing all the data sets available, both singly and in combination.

What this evolution reflects is the understanding that the results of a survey are not just a stand-alone set of tables, but an addition to an information base that may be used in many foreseen and unforeseen ways.

By entitling this Symposium "From Data to Information" we aim to focus attention on problems and issues faced in the process of ensuring that our data, our numbers, become information by adding to knowledge. The challenge is to find the methods and solutions that will enable statistical agencies to ensure that their valuable data assets become information that is both useful and used.

This will take us into many domains which I trust are well represented in the programme of this Symposium. We will cover information management and data warehousing - how we organize and make available our data holdings and information about them; we will cover data integration and how we can facilitate the joint use of data collected from different surveys; we will cover issues of access to data for analysts and how

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26-J, R.H. Coats Bldg., Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

we can enable them to perform the analysis they want while preserving the confidentiality of individual records; we will cover the dissemination of information of data to the general public and the role of the media; we will cover issues of data quality and analytic methods that take account of both the data source and the data quality; we will address the impact of technology on dissemination; and we will address how partnerships can facilitate the role of statistical agencies within the information industry.

Why are we addressing this subject now? At least four reasons that I can think of.

(a) More emphasis is being put on understanding processes and factors that influence them, rather than just describing the current state of affairs which result from these processes. We want to understand factors associated with people moving into and out of poverty, for example, not just how many do. We want to understand why some businesses succeed and others fail, not just how many.

(b) I already mentioned financial pressures. Tight budgets and the high cost of data collection dictate that we fully utilize our existing data holdings. This includes not just the full exploitation of single data sets, but also the integration or matching of different data sources.

(c) Technology makes new things possible. We can deal with more data faster and easier. We can disseminate data more widely and faster. We can consider computer-intensive methods of analysis that would have been out of the question a decade ago. On the other hand technology comes with a price. We are faced with tougher issues of data management and control, of allowing access when appropriate and preventing it when not. And we must face the privacy issues associated with linkage and with secondary use of data.

(d) Finally, rich databases, especially longitudinal databases, make richer analyses feasible. The potential for extracting greater information and understanding is itself becoming greater as we accumulate larger and richer databases whether from surveys or administrative sources.

These are some of the motivations for pursuing this topic now. At Statistics Canada we are facing many challenging problems in this domain. It is gratifying to see so many people have come, many from long distances, to share their experience and expertise in helping us resolve these problems. I am very pleased to see that we have here on the programme of this Symposium many very distinguished speakers in areas pertinent to our topic.

You may be interested to know that we have somewhere between 250 and 300 participants in this Symposium, including representatives from many statistical agencies, from other federal departments in Canada, from provincial governments, from universities, and from the private sector.

We have a lot to cover and I should not delay proceedings further. So let me thank you all for coming to participate in this Symposium. We are looking forward to a very fruitful three days. I know that we at Statistics Canada will benefit from the presentations and exchanges, but I hope that everyone will be able to go home with some new ideas, new solutions, or at least new information that will benefit them and their organizations.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

THE ROLE OF STATISTICS IN MAKING SOCIAL POLICY

P. Hicks[1]

The theme of this conference is "From Data to Information". Information is not an end in itself. When information is used, it becomes knowledge. My remarks in opening this conference are intended to set the stage by addressing the broader question "From Data to Information to Knowledge". In particular, I will talk about how statistical knowledge is used to make social policy.

I will conclude that fiscal and governance pressures are leading to a more empirical social policy. The statistics are *not* now in place to support a social policy that is driven by evidence.

However, Canada is in an excellent position to become a world leader in building the information base to support a new, results-oriented social policy. The result will be savings in cost, dramatic gains in effectiveness and more empirical approaches to governance and national standards.

DEFINITIONS AND ASSUMPTIONS

Let me begin with some definitions.

By policy, I refer to the activities of those who advise governments about the direction that programming should take.

Hard data and soft data. The information that supports policy can be hard or soft. Hard information includes the familiar national surveys by Statistics Canada. It also includes the kind of information that one finds in good program evaluations – what happened to groups of program participants before, during and after program interventions and how that compares with control groups.

The other type of information is soft information – information on values and attitudes. This is the information that comes from public consultation and polling. At what point do the public and experts think that a social trend, like that relating to family violence, becomes a "policy problem" that warrants priority government action? What do public opinion and the advice of experts tell us about appropriate programming for young offenders?

Without good hard data, we must rely on intuition, anecdote and ideology. Without good soft data, policy will get out of tune with politics and reality.

In this talk I will limit myself to comments on the state of hard statistical data. However, it is important to draw attention to soft data, since it is as important to the policy process as hard data. And, in my view, it is equally in need of major change. However, that is the subject for a different talk.

A statistical field of dreams. Before turning to an assessment of the state of social statistics, I would like to make explicit an important assumption. It is that, if there are good statistics, they will be used in the policy process.

Many in this room may be sceptical. There are countless stories about there being an information overload and that much existing data simply sits on the shelf, unused. Often statistics are seen as justifying decisions that have already been taken, not as an integral part of policy formation.

My experience is the opposite. There is a real hunger for relevant statistics. If they are not used, the conclusion I draw is that they are not relevant. People want to do things right and to do the right things. They welcome evidence that will help in this. There is no one advocating employment policies that don't work, or health care interventions that make people sicker.

Often there will be some temporary self-interest in maintaining policies that don't work – like professionals whose skills will become redundant because of reform. However, even here, public pressure and a sense of professionalism will ensure that this resistance is only temporary, if there is good evidence that shows what

[1] Peter Hicks, Senior Policy Advisor, Government of Canada, 56 Sparks Street, Room 600, Ottawa, Ontario, K1P 5A9.

would work better.

Statistics must be known before they can be used. There is a complicated route to follow before statistical information is actually used in a policy decision. Cabinet meetings are not devoted to studying the latest tables from Statistics Canada. Rather, statistical information is used to inform the views of those advocating change, and to help policy advisors assess the different options for change that are eventually presented to ministers.

There are weak links along this route. Policy departments in Ottawa and provincial capitals are often driven by the problem of the hour. Many have only a modest capacity to deal with longer-term issues in a quantitative way. Independent policy think tanks, as well, often have short time horizons. Interest groups are often too fragmented to take an empirical perspective. And there are few academics who have the time or resources to devote to this interface between statistics and social policy.

Nevertheless, my contention is that these links would be established quite quickly and easily, if the relevant statistical data were there – that is, a statistical field of dreams.

The remainder of this talk tries to answer three questions. First, what would an ideal set of social statistics look like? Second, how close are we to this ideal? Third, how can we improve things?

IDEAL SOCIAL STATISTICS

The ideal set of social statistics would be comprehensive and integrated, both in terms of breadth and depth.

The horizontal dimension. By breadth, I refer to detailed data that can be compared across the different domains of social policy. The old boundaries are eroding quickly. As recently as a generation ago, income support and social services were thought of as quite separate departments of social policy. The worlds of school, work and retirement were routinely dealt with in separate boxes. The obvious links among poverty, health, the nature of work, social integration, learning, crime and well-being were not incorporated into social policies.

Today all this is beginning to change. The idea that everything is interlinked is now well established, at least at the level of rhetoric and keynote speeches. It is beginning to be felt at the operational level as well. For example, the recent creation of the federal department of Human Resource Development brought together social programming that was formerly thought to be quite separate.

An ideal set of statistics would support this new, more comprehensive social policy. It would shed light on what happens in homes, schools and doctor's offices. It would deal with intergenerational and life cycle issues, the changing nature of work and learning, and the combined impact of programs on individual and societal well-being and development.

The vertical dimension. By depth, I refer to an integrated data base that supports many kinds of users. The same data base should provide social indicators that are used to signal new problems and mobilize public action. It should support program evaluators who ask what has worked in the past, and policy advisors who ask what might work in the future. It should support front line staff of service agencies who ask what particular intervention will work best in which circumstance. And it should support individual Canadians who want to know which kind of training courses, or health interventions or exercise regimes are most likely to work best for them.

The ideal data would therefore be strong enough to measure the effectiveness of programs – costs and results or outcomes – and to predict which interventions are likely to be successful.

The ideal data base would allow this program-oriented data to be integrated with information about the functioning of all the major institutions of society – the family, the labour market, the community, the schools, institutions of health care and social services, and cultural organizations. It would measure trends in individual activities and individual well-being.

TODAY'S REALITY

The second question is how far are we from this ideal.

A full, qualified answer would point to many recent improvements and would stress that even more improvement is at hand. It would point out that Canada is in much better statistical shape than other countries. It would point to the excellence of existing data for use in marketing applications and to the power of the new longitudinal surveys for research purposes. This would include the powerful new SLID, health and childrens' surveys.

Nevertheless a poor report card from a policy perspective . . . However, the short answer is that – from the perspective of the social policy maker – the situation is poor. From the perspective of policy makers, the only thing that really counts is consistent, integrated statistics that track problems, and potential solutions, over time. Here there has been little change in the basic statistics for decades.

A digression on the importance of trend indicators. Let me digress for a minute on why policy is so dependent on good indicators of trends.

> Academic researchers are interested in the causes of problems – and the in-depth, often longitudinal, studies that provide this information. Marketers are interested in data with rich cross-classifications, by geographic area or socioeconomic grouping.
>
> Policy advisors find this kind of information nice to have as well. However, for policy purposes, it is long, consistent time series that are important. They must cover many years, so that the cyclical dimensions to problematic situations can be understood, as can the impact of different government programs.
>
> Why is knowing about trends so important? We are a society with highly developed social institutions that already address just about any conceivable social problem. The inevitable choice confronting governments is to do a little more or less of something, or to do things in a slightly different way, or to shift responsibilities to or from some other order of government or the private sector. What drives these decisions is information on whether problematic situations are getting worse over time, or whether government programs are becoming more or less effective, or affordable.

In the areas of greatest interest for social policy today, we have virtually no good trend data, although the general social survey and some labour force survey supplementary surveys provide a few glimmers of light.

. . . but much that is still worth preserving. We do have good statistical time series to support the policy issues that were important in postwar period up to about the end of the end of the 1960s – when social policy meant the provision of a strong social safety net, addressing income inequalities that arose from the operation of the market, and providing wide access to institutions like schools or hospitals or paid work.

That is, we do have reasonably good information on trends in the incomes of Canadians, on who participates in what kind of institution and on the associated costs. These are still necessary things to know, even if the main focus of policy has shifted to other topics.

In other words, I am not suggesting that we drop the statistical data that we now have. They are still relevant for many issues and the advantages of having consistent, long time series are enormous even if the conceptual framework underlying those time series is outdated. Rather, the simple point I am making is that the existing series do not stand up well when viewed from the perspective of current policy issues.

LOOKING TO THE FUTURE

The final part to my talk is addressed to what is to be done. I have several observations to make – all optimistic in tone.

In summary, we are close to having a framework that will allow the needed integration. The technology is in place to handle the huge of amounts of data in question, without infringing on privacy. And there are strong fiscal and governance reasons for change.

A framework. I have described an ideal set of social statistics that is very comprehensive. At present many of the pieces are there, but there is no framework that would allow them to be used in an integrated way. Fortunately, we are very close to a new conceptual framework that can underpin both social policy and social statistics.

On the policy side, we are either too shell-shocked or otherwise not brave enough to talk explicitly about big frameworks. However, despite our timidity, a framework to guide policies is nevertheless emerging. The OECD has set its new directions in social policy. Similar concepts are increasingly used in all social disciplines and in all developed countries. These centre around ideas like human development, investment in human capital, lifelong learning and other life-cycle approaches to policy, far more attention to family and inter-generational issues, and thinking of work in terms of skills rather than occupations.

I believe it will not be long before ideas like these will result in a statistical framework that will do for social policy what the system of national accounts does for economic policy. It will be an even more powerful than the national accounts in that it will be based on how people spend their time, an even more comprehensive and fundamental concept than how people spend their money.

The new statistical framework would account for how people spend their time in schools, at work and at leisure – with what degree of social interaction and under what constraints, with how much learning and earning, and how much satisfaction. The new framework would allow us to keep track of changes in how Canadians spend their lifetime hours in school, at home, at work, in retirement, in giving care to others, in receipt of various kinds of government programs or associated with various institutions.

Technology. The technology is now in place to manipulate the huge amounts of data required by this framework. Many speakers at this conference will discuss technology and I will not dwell on it here.

The result is quite simple to describe – a set of interlinked micro-simulation models fed by existing and new surveys, and by administrative records. Huge amounts of information will be stored about synthetic individuals and their institutional attachments. Privacy is not an issue because the "individuals" in question are not real Canadians. Rather the data will be based on imaginary people who, when taken as a whole (or in groups), have the same characteristics as real Canadians (or sub-populations of Canadians).

At this point, I expect the eyes of many in the audience will have begun to glaze. Huge data bases? Integration across different areas of social policy? Data on what actually works? Imaginary people? No threat to privacy? Dramatic increases in effectiveness? This may not sound a lot like the world we now live in.

Remember that we now see the world through statistical glasses that were designed a long time ago. The system of national accounts, the census and the labour force survey were all designed before the computer age. The power of these instruments has been greatly expanded by computers, but the underlying structure is based on the technology and paradigm of the adding machine. It should not be a surprise that dramatic improvement will be possible when designs are built completely anew around a powerful new informatics technology.

There is a plan. I would like to assure you that what I have described is not a day dream. A plan to do all of this now exists.

. . . that will shift programming to an empirical basis. In terms of the program-oriented data, Human Resource Development Canada is now testing a model that will simultaneously allow policy people to assess the cost effectiveness of different kinds of policy interventions and that will allow individual Canadians to have data on the probability of success, for them, of different kinds of training or other employment interventions. The system is based around a data warehouse that will allow current decisions to be assessed against the success of all past interventions for people with similar characteristics and in similar circumstances.

This kind of technology will revolutionize social service programming and allow it to move to an evidence-driven basis. It will shift decision-making to the front-line of service-providing organizations. They will become true learning organizations that continuously adjust in response to new evidence. Eventually, the new technology will put information and power in the hands of citizens themselves. The savings to government, as this technology spreads across the various social and health disciplines, will eventually be in the billions. The contract between citizen and state will be rewritten. Social programs will actually work.

. . . and revolutionize our knowledge of society. Similarly, with respect to the statistics about the population as a whole, a draft proposal has already been developed by officials at Statistics Canada and Human Resource Development Canada. It describes the new framework, the new micro-simulation models and the new data collection that is needed. Its implementation would, as well, revolutionize social statistics and social policy – that is, our understanding of society and how to make things better.

This statistical proposal has not yet been funded or validated by a wider community of interest groups and academics. Funding is always tough, especially in a time of restraint. However, even here, I am guardedly optimistic. There are powerful pressures that will support moving social policy towards a more empirical basis, even if it involves up-front funding. These pressures are related to the fiscal situation and to governance.

Fiscal pressures. Fiscal pressures will result in a new attention to effectiveness, doing things that actually have a payoff to individuals, to society and to government treasuries. This has not happened yet. But once the various rounds of indiscriminate initial cuts to social programs are over, attention will almost certainly to shift to an examination of which remaining program elements are actually working and which are not.

Governance pressures. Governance would be so much easier with a more empirical approach to social policy.

Today, for example, when we think of the big issues in social policy, we typically think about national standards or constraints in areas like access, or financing or processes. We are concerned that programs be portable, or without user fees, or that only certain public sector or certified bodies can deliver certain services.

Here we are in the world of huge fiscal transfers, abstract principles and big constitutional arrangements.

Tomorrow, we will be more concerned about standards as they relate to results or outcomes. Attention will focus on comparability in different parts of the country in the usefulness of skills obtained, not just the pedagogical standards in various training programs. The focus will be more on the comparability of health outcomes, and less on accessibility issues like which health care services are deemed to be medically-necessary.

We will be in a world where the main costs associated with national standards and principles will be the relatively minor costs of surveys and research, not fiscal arrangements.

We will be in a world where standards can be dealt with, not in terms of abstract societal values, but in terms of the more prosaic, but more real, standards that are embedded in the design of statistical indicators.

We will be in a world of decentralized decision-making, with many partners consulting over empirical issues – users, practitioners, academics, departments of governments at all levels, even international organizations. As the boundaries of social policy crumble, and as social policy gets increasingly inter-twined with tax policy and economic policy, many players must necessarily be involved in ways that are increasingly interrelated.

To be clear, I am *not* suggesting that standards or principles relating to inputs or process will no longer be needed. Those that deal with mobility and portability of benefits will continue to be particularly important.

I am *not* suggesting that fiscal transfers are unimportant. They are needed to ensure comparability in the capacity of provinces to finance good social programs.

I am certainly not downplaying the importance of constitutional solutions – the sorting out roles and responsibilities and the elimination of overlaps and gaps – especially now in post-referendum Canada.

What I *am* suggesting is that, from the perspective of social policy, the fundamental challenge should be to find ways that will allow the many partners, who must necessarily be involved, to work together productively. Good statistics will provide the common language that will make this possible.

A common statistical language will make it much easier to deal with all social policy issues, including common standards and goals. It would allow substantive progress to take place at a decentralized, results-oriented level, based on evidence.

Higher level discussions – whether in the context of constitutional arrangements or social charters or the principles to be associated with the new Canada Health and Social Transfer – could be particularly productive if they signalled a shift to this more empirical approach.

CANADA CAN LEAD

Canada is an ideal position to play a world leadership role. Statistics Canada, working with its many partners, it is very well placed to take the lead in developing the new common language. It is already a world leader in the area.

With strong leadership, and much persistence, we can lay a foundation that will enable our successors to build a better society in the next millennium.

We can make a big difference.

SESSION 1

Data Integration

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

DATA INTEGRATION: THE VIEW FROM THE BACK OF THE BUS

G. Priest[1]

ABSTRACT

Statistical agencies have tended to be methods driven. That is, collection activities took place through vehicles developed around specific methodologies. Each vehicle often served its own specialized clientele without regard to the needs of other organizations. The agency, therefore, often evolved, not as a corporation but a consortium, or even fragmented consortium, of relatively independent producers of data. Methods, systems, concepts, definitions, classifications, products and services were developed independently resulting in inefficiencies, redundancies, disharmonies and some client frustration.

The client satisfied with single-source information has been relatively well-served. But the client who needed comprehensive information on a particular issue, population or geography has not. Now information technology has precipitated a paradigm shift.

A new generation of clients is cutting its teeth on the Net and developing new expectations, particularly with respect to searches for information. These clients, all with their unique and particular needs, expect to be able to thematically browse meta information, determine sources, make selections and even download: on-line, real-time, seamlessly and at low or no cost.

The challenge to, and opportunity for, the statistical agencies is to respond to the new paradigm by accommodating these clients. The keystone to building such a response capability rests in integration. This includes both developing links between the sources and eliminating or reducing the disharmonies. Integration is also fundamental in moving from data to information because it facilitates bringing together all relevant and available inputs. Informed decision making depends on it.

KEY WORDS: Silos; Meta information; Disharmonies; Single-source output; Thematic; Integration.

1. INTRODUCTION

1.1 Integration

In the early 1970s I attended a meeting of the Conference of European Statisticians, the purpose of which was to develop a harmonized set of international data on housing, households and families. At dinner, after the first day of meetings, another Canadian delegate who was from Canada Mortgage and Housing Corporation, asked how I could be such a staunch advocate of the harmonization initiative. Surprised, I asked him what he meant since it was quite obvious what needed to be done. “I make the point”, he replied, “because your own department is in such disarray”. He noted that he had to go to as many as eight divisions in Statistics Canada to get the data he needed. Furthermore, he often found that there was a lack of comparability between sources. I immediately became an advocate for integration but to little avail. There was little interest in collaboration between my colleagues in the department.

In the mid 1980s my division launched two products which attempted to integrate data from diverse sources. One was the quarterly **Canadian Social Trends**. The other was the **Target Group** series of publications. In these ventures we became users ourselves and suffered many of the frustrations of external clients. The task of searching for, and gaining access to, the data we needed was formidable. We often likened our situation to that of the struggle for racial integration in that we felt very

[1] Gordon Priest, Director, Integration and Development of Social Statistics, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

much that we had been relegated to the back of the bus and we asked how this could have happened.

The situation had evolved because most statistical agencies are organized on the basis of differing methodologies. We have censuses, post-censal surveys, household surveys, business surveys and the derivation of data from various administrative records. Thus, our data gathering has tended to be vehicle-driven. And each vehicle tended to develop its own expertise in systems, methodology and subject-matter.

1.2 Islands, silos and stove pipes

This is what Tapscott and Caston (**Paradigm Shift: The New Promise of Information Technology**, New York, McGraw-Hill, 1994) have referred to as "the problem of the unregulated enterprise". They describe islands of technology or expertise which meet specific needs but result in a fragmentation of the organization. They note that such islands have limited and specialized functions that may have nothing to do with overall business objectives or strategies of the corporation. Furthermore, they may become balkanised with formidable physical and organizational barriers, redundancies and inefficiencies. They cite lack of integration as a source of significant loss in business opportunities.

Keith Vozel of AT&T , in his "Technical Evolution White Paper", described such organizations as vertical or stove-pipe, the parts of which tend to address a single issue or client without regards to the needs or requirements of others. These organizations are wasteful in terms of redundant or replicated data in which there is no enterprise or corporate view of the holdings. Other literature refers to such organizations as silos to which access is difficult and between which communication is non-existent or limited. They represent untapped potential and lost opportunities.

1.3 The Consortium and its implications

Bill Bradley of Health Canada has described Statistics Canada as just such an organization of autonomous data development programs. My own view of statistical organizations is that they are less corporations than they are consortiums of independent producers. While many of these producers have well-served their specific clients, it has not been without a price.

2. DISCUSSION

2.1 Lack of meta information

Statistical agencies generally have very little *corporate* knowledge, if any, regarding the nature and extent of their data holdings and what knowledge they do possess, has not been systematically shared with clients and potential clients. How often have we heard a policy maker, decision maker or researcher lamenting the lack of data when suitable data actually existed but were buried away from sight in some antiseptic and air conditioned tape library? Unfortunately, the production of meta information (that is, information about the data holdings), is very dependent upon the various production areas. The amount of meta information that is held may vary significantly from area to area and it is not usually documented to any corporate standard. Where attempts have been made to develop standardized meta information it is more likely to serve some bureaucratic purpose rather than potential clients. This results in underutilisation of the data collections. Clients, as well as agency staff, undertaking research on any given issue or population, are left largely to their own devices to contact *each* of the source areas to determine if they have any relevant data. Speaking from experience, I can say that the task is formidable, frustrating and often, fruitless.

2.2 Disharmonies

As might be expected, given the nature of independent production, further complications exist due to disharmonies between vehicles or sources in terms of concepts, definitions, classification systems, and documentation. Not only has each production area developed its own methodological, processing and dissemination practices, so has it developed its own subject-matter content. Through lack of care, concern or communication, differences have arisen in terms of concepts, definitions, classification systems and database coding. Not only is this distressing to the end user but it is also wasteful of resources. Given the lack of corporate standards, program managers, time and again, have developed totally new documentation, unmindful of what might already have been produced elsewhere in the agency.

I am sure we are all aware of those situations where a data set from one source cannot be compared with another source, even though it bears the same name. On the other hand, there are those cases where variables actually are comparable but carry different names. At Statistics Canada we have even uncovered situations where variable names may be comparable in one official language but not in the other. And we have probably all experienced those situations where, even though a variable may carry its conceptual integrity from one source to another, comparability may be lost because each source used a different classification system or used non-standardized aggregations. Finally, there are those insidious practices of using different mnemonics in the coding of variables on record layouts for micro data retrieval. This can lead to serious coding errors for persons working with multi-source files.

2.3 Contradictory or incomplete outputs

Another legacy of our stove-pipe production is that of independent vehicle-driven output. There are the obvious difficulties when Survey B contradicts the earlier released figures from Survey A indicating that the annual number of Peggy's Cove tourists hit by seagull droppings is 5349 not 316. Such incidents are followed by the usual flurry of releases containing footnotes and qualifications explaining that one source was seasonally adjusted, included great blue herons with the offending seagulls or was rounded to prevent residual disclosure. Or sometimes we just issue our blushing pink errata sheets and *'fess up'* to a "computer error". While these situations are embarrassing they do not often cause long-term damage because they are relatively rare and usually quickly identified and corrected.

2.4 Single-source output biased

Of greater concern is the analytical output that releases a set of information from a single source without the benefit of related and relevant data from other existing sources. Such releases can be dangerous in terms of providing partial and therefore, biased and misleading information. That is, the information is not set in the context of our comprehensive knowledge of a situation. As an example, suppose a survey of the social drinking of young women reveals that one out of twenty of the population surveyed had been assaulted by a young male after leaving a bar late at night. The information is released, there is great public discussion and a demand for legislation to close bars earlier. Suppose other information existed from a comprehensive survey of violence in society: including violence in the home, in the workplace and in the street. Suppose that other survey revealed that, while the social drinking data were substantiated, an even higher rate of young men were assaulted when leaving bars late at night. Suppose it also revealed that young men were the greatest offenders in terms of assaults, not only in the streets, but in the home and the workplace as well. The ensuing public discussion and search for a solution would probably have been substantially different if the information from the drinking survey had been placed in the larger context.

2.5 Implications of stove-pipe production

To summarize the implications of stove-pipe production in statistical agencies, we see that corporate knowledge of the extent and nature of their data holdings may be incomplete and therefore, of diminished use to the client. Disharmonies exist between sources and, therefore, even when the client does find different sources of interest, the data may not be comparable. Finally, the agency may mislead clients by releasing vehicle-driven data rather than integrated outputs. If we accept that fragmented production poses a problem for clients then we have to consider integration as a solution. That is, we need a corporate inventory of our holdings, we need to resolve the disharmonies and we need to ensure that data releases are made in the context of our full knowledge of a situation.

2.6 Compelling reasons for action

There are compelling reasons to take these actions now. Firstly, many agencies are faced with funding cuts at a time when the demand for information is increasing. It is understandable that in tough economic times policy makers and decision makers in both the public and private sectors want the most reliable, most recent data because the implications of making a wrong or uninformed decision are far more serious. It falls, therefore, to the statistical agency to not only do more with less, but to work smarter and that includes mining and utilizing existing data as fully as possible. And you can't mine what you don't know you have. Maintaining dynamic corporate metadata and information just makes good business sense.

Secondly, technology now exists to make the job of data and metadata management infinitely easier than was the case ten or even five years ago. In 1980, Canada Mortgage and Housing Corporation asked if I would estimate the cost of building and maintaining a meta information base for housing. We estimated that it was going to cost over three person-years to build it and about one-and-a-half to maintain it. Needless to say, the meta information base was not built. This year we built a meta information base for all social statistics which listed approximately 20 subjects, over 1000 variables and close to 100 sources. It cost us less than half a person-year to build and maintenance will be negligible. It provides the client with the option of using both thematic and keyword searches. Using the thematic approach, the client browses the list of subjects or themes (e.g., demography, education, ethno-cultural, health, labour, etc.). Selecting a subject reveals an alphabetic listing of all associated variables. All sources for each of the variables is shown. Selecting a source of interest reveals baseline information about the source, a further thematic list of variables for that source, micro data record layouts, questionnaires and other documentation. Advantages of the approach are that the client becomes aware of nests of related variables that may prove useful and variables may be cross-linked to a number of different subjects.

Thirdly, clients, especially those with Internet experience, have become increasingly knowledgeable and sophisticated with respect to searching for information. Thus they have increasing expectations of being able to approach a statistical agency, browse its holdings, specify output and download it; online, real-time at low cost or no cost. While there will be undoubted costs in building such a service capacity there is also a potential for hard cost reduction (cost avoidance) and improved productivity. For example, agencies should reduce the number of expensive generic products and allow and encourage or assist clients to build their own niche products.

3. FUTURE ACTIONS

3.1 The vision

Thus, there is need and there is opportunity. We must develop the vision and the corporate will to accept the challenge and seize the opportunity. There are three fundamental components to the vision. Build the meta information, resolve the disharmonies and move from vehicle-driven outputs to issue (or population)-driven integrated outputs.

3.2 Building the meta information

Meta information must be comprehensive. It must respond equally to the client who simply wants an answer to a question such as the number of widgets produced last year as well as to the client who wants to know what is resident on micro data bases so he or she can do his or her own research. Therefore, meta information must describe the contents of micro data files, the contents of aggregated tabular output, the content of analytical or descriptive reports and the nature of specialized services provided by the agency. The information must be accessible by a search tool that facilitates both keyword and thematic searches. Ideally, a thesaurus should sit in front of such a tool to translate the client's lexicon to the agency's lexicon. The importance of a thematic search tool cannot be underestimated as is witnessed by many of the more helpful sites on the Net. The listing of subjects or themes and variables associated with those themes enhances the search by revealing variables that may be useful but not previously evident to the client. Regardless of whether the client searches on the basis of keyword or themes, however, the outcome should be the same. That is, he or she must be directed to the *source* of the information or data sought.

3.3 One gateway: one tool

Experience has shown that clients have found the statistical agency to be a bewildering maze of seemingly illogical sources. I can attest to this in the many calls I have received over the years which were prefaced by, "I don't know if I called the right place, but do you have...?" There must be one gateway to the organization and at the gateway must reside *one*, user-friendly tool, or knowledgeable helpful staff equipped with the tool, capable of directing the client to the appropriate sources. Different systems might underlie the one tool as long as a common look and feel is maintained.

The gateway may be replicated at different physical sites, but, again, it must have the same look and feel at each. It may be electronic and fully automated or supported by advisory staff. With regard to an Internet site, caution must be exercised with regard to channeling the entrepreneurial spirit and constraining the egos which have seen "home pages" blossom as the vanity press of the electronic media. Each such initiative should be questioned in terms of what it costs to build and maintain and how effectively it contributes to the client's search. We must avoid the pitfall of building stove-pipe solutions to stove-pipe problems.

In the course of this symposium you will hear at least two presentations from my colleagues at Statistics Canada which represent parts of an ultimate solution. One is the IPS system, the other is Statcan On Line. Each represents part of a solution but they must be integrated at some point into part of a single corporate strategy.

3.4 On-line, real-time

In a very short period of time the Net has significantly raised our expectations in our quest for information. We are satisfied with nothing less than instant, electronic gratification. While the Net is perfectly positioned to assist the client in browsing our meta information, the question arises as to how we deliver a real product or service when the client finds something he or she wants. Clients now are less satisfied with generic products as we have seen the evolution of niche markets in which clients demand custom output suited specifically to their needs.

Once a client has been directed to a source of interest, it is in the client's interest and the agency's interest to provide the client with the facility to download, on-line in real-time that information or data sought. The client's interest is obvious but the agency's interest is served in not only happy clients but also in hard cost reduction. The greater the capacity for a client to browse, specify, code or download, the less resources consumed by the agency. The technology exists to allow clients to download from public use micro data files and be billed automatically. Only in the case of confidential master retrieval files (which must remain behind firewalls and screened for residual disclosure) is there a need to distance the client from the data. But even then, there is no reason why the client cannot code the request from record layouts, submit the job and have the agency produce the output and do the necessary disclosure screening.

The issue of billing and costs to clients, while a fascinating subject, is well beyond the scope of this paper and this symposium.

With regard to the client who does not have the skill or the time to down-load his or her own data and information the option should be provided for account executives, using the same tools, to custom-build outputs to meet client's niche needs. As the meta information opens the data archives to the world it might also be expected that opportunities will develop for private sector consultants to undertake browsing, downloading and analysis on behalf of clients.

3.5 Addressing the disharmonies

It is unrealistic to think that all disharmonies can be eliminated between sources. Differences in methodology such as whether a question is asked on the doorstep, over the telephone or on a self-completed form may yield subtle differences in output. Nevertheless, most serious disharmonies can be eliminated with concerted effort. I was associated with such an effort a few years ago to bring harmony to family data from some nine or ten sources. All of the serious and most of the minor disharmonies were eliminated by negotiation between the production areas. It is not, however, a one-time effort. As new sources came on-line new disharmonies developed. One of the most formidable tasks in the exercise in question was simply identifying all the sources of family data. We had, in effect, to build an inventory before we were able to identify and address the disharmonies. In that regard, the building of the meta information facilitates the identification of the disharmonies. In our recent undertaking of building meta information on social statistics many disharmonies were revealed in the process. Each was flagged for future attention. The meta information can also become a model of best practices and even a template for the development of standardized documentation ranging from mnemonics used in record layouts to classification systems to definitions. The adoption of templates and standards also promises the potential of hard cost reduction as future sources are developed. There is, however, no avoidance of the discussion and negotiation that must take place between the source areas with a view to the development of those standards. And there must be a commitment to eliminate the disharmonies.

3.6 Increased thematic output

The integration of data in a thematic way will also be facilitated by the construction of meta information. In the past, analysts may not have known of many relevant sources which existed, but armed with appropriate meta information, search tools and retrieval systems there is no reason why all relevant data cannot be ported to the desk-top. It remains, however, for the analyst to understand the importance of integration. At least, aggregated or tabular output should be accompanied with pointers to other related sources. At best, analytical or descriptive output should incorporate *all* relevant data and information in the analysis or discussion. It must be realized that the release of anything less than our comprehensive knowledge of an issue or population is as potentially damaging to our clients as are undetected response or processing errors. It is indeed curious that the statistician who shows such a proclivity for footnotes on methodological issues should have been so silent with regard to other sources of information or data relevant to the client.

3.7 Corporate Initiative

The question remains whether the above-noted steps can be undertaken without corporate initiative. As long as the corporate culture is such that it rewards individual production rather than corporate production it is doubtful that change will happen. Unless the stove-pipe production areas perceive some advantage in improving whatever performance measures against which they are evaluated they are unlikely to take initiatives. Perhaps some will, creating a groundswell in which others must join or be left behind. Even so, is there not too much at stake to leave such developments to random individual acts? Is there not the possibility of duplicated effort and wasted resources? Does the lack of a shared vision, strategic planning, direction and funding from the corporation send the signal that integration is not really a high and urgent priority?

Information technology today presents unique challenges and opportunities to statistical agencies but to seize them it will be necessary to place a high priority on integration. That suggests the establishment and funding of a centralized body within the organization charged with leading the above-noted activities.

4. CONCLUSION

4.1 The past

The organization of statistical information has been driven primarily by methodology rather than thematic content. The integration of data on the basis of issues, populations or geography, and attempts to convert those data to information, have been hindered by the structure of the silos in which they have been collected and archived. There has not been a corporate, or for that matter, client view of the richness and comprehensiveness of the data holdings.

4.2 The future

In the statistician's ideal world there would probably be complete record linkage between all sources of data and, as a result, full integration. Few, if any, agencies, however, operate in societies that would tolerate such a manipulation of private information. The challenge, and the opportunity, therefore, lies in moving to corporate rather than consortium data management. Meta information, harmonization and thematic integration are imperative if we are to progress in moving data to information. Agencies which fail to accept the challenge and the opportunity provided by information technology and who continue to relegate their clients to the back of the bus, particularly clients who have traveled the information highway and like what they have seen, will be quickly perceived as unhelpful and irrelevant.

5. REFERENCES

Bradley, B. (1994). Metadata matters: Standardizing metadata for improved management and delivery in national information systems, discussion paper, Ottawa, Health Canada.

Hammer, M., and Champy, J. (1993). *Reengineering the Corporation*. New York: Harper-Collins.

Nordbotten, S. (1993). Unpublished paper. The Statistical Meta Information System Workshop, Luxembourg: Eurostat.

Probst, S. (1995). Keynote Address, Data Warehouse Symposium, Ottawa: Tanning Technology Corporation.

Tapscott, D., and Caston, A. (1994). *Paradigm Shift: The New Promise of Information Technology*, New York: McGraw-Hill.

Vozel, K. (1993). Technical Evolution White Paper, discussion paper, New York: AT&T.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

META-ANALYSIS OF MULTIPLE COHORTS OF UNDERGROUND MINERS EXPOSED TO RADON

Y. Wang[1], D. Krewski[1,2], J.H. Lubin[3] and J.M. Zielinski[1]

ABSTRACT

Methods of meta-analysis of a series of cohort studies by using random-effects models are presented in this article. Specifically, a non-linear random-effects regression model was used to describe both population average and cohort specific risks. The methods were used to fit proportional relative risk model for estimating dose-response relationship of radon exposure and lung cancer mortality. Because of the computational burden involved in fitting the non-linear random effects model, a two-stage regression approach to meta-analysis is also considered. The results obtained with the random-effects and two-stage methods are compared with those based on conventional meta-analytic methods in which overall estimates of risk are based on a weighted linear combination of cohort specific risks (with weights inversely proportional to the precision of the estimates).

KEY WORDS: Cancer mortality; Cohort study; Generalized estimating equations; Lung cancer; Radon progeny; Random-effect model; Heterogeneity.

1. INTRODUCTION

Meta-analysis of occupational epidemiology studies. Meta-analysis is a quantitative method of data aggregation (Greenland, 1994). It is used to identify overall effects for all studies combined, and to characterize differences among individual studies. Conventional meta-analysis methods effectively average the outcomes of available studies with the weights assigned to individual estimates being inversely proportional to estimation errors (Greenland *et al.*, 1992). Currently there is a trend towards the use of random-effects models in meta-analysis (Berlin *et al.*, 1993; Berkey *et al.*, 1995). The National Research Council (1992) recommends the use of random-effects approaches to meta-analysis and the exploration of sources of variation among study results. The advantage of random-effects analysis over conventional meta-analysis techniques is some allowance for sources of heterogeneity beyond sampling error (Greenland, 1994). Random-effects analysis provides an estimate of effect aggregated across all studies, as well as estimates for individual studies. The overall estimate is referred to as a fixed-effect whereas the study specific estimates are called random-effects (Moolgavkar *et al.*, 1995).

In this article, we will explore the use of a random-effects model in meta-analysis of occupational epidemiology studies. Our motivation is a need to synthesis relationships of radon exposure and lung cancer mortality from 11 major underground miners' studies conducted in Canada, the United States, and the others countries, for which there is significant heterogeneity between those studies (Lubin *et al.*, 1994). Radon is an inert gas produced during the radioactive decay of uranium. Alpha particles emitted during the radioactive decay of its short-lived progeny are responsible for the carcinogenic activity of radon. A meta-analysis of underground miners studies conducted by Lubin *et al.*, (1994) has clearly established that radon decay products are carcinogenic and that exposure to radon progeny at levels found in miners increases lung cancer risk. The characteristics of these studies are

[1] Health Protection Branch, Health Canada, Ottawa, Ontario, Canada.

[2] Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

[3] Biostatistics Branch, National Cancer Institute, Bethesda, Maryland, U.S.A.

briefly summarized in Table 1. Although significant increasing trends in lung cancer risk with increasing exposure to radon were apparent in each of these studies, estimates of the excess relative risk per working level month (ERR/WLM) exposure to radon derived from different cohorts varied substantially. Therefore, the between studies heterogeneity needs to be considered.

In this article, we explore methods of meta-analysis of cohort studies. In section 2, random-effects analysis methods are presented. In section 3, two-stage analysis methods are discussed. An illustrative application of these methods is given in section 4. Our conclusion are given in section 5.

2. STATISTICAL MODELS

Poisson regression methods are often used in analyzing data from cohort mortality studies (Breslow and Day, 1987). With this approach death rates are assumed to be constant within fixed time intervals and exposure categories. Data entering into regression analyses are in the form of a multi-way person-year table consisting of the number of deaths from disease of interest and person-years, of observation, classified in terms of relevant covariates. Under the Poisson regression model, the observed number of cases is assumed to follow a Poisson distribution for which the variance is equal to the mean. Specifically, the expected number of cases is modeled as

$$N_{jk}\, r_{jk}\,(x,\, v), \tag{1}$$

where N_{jk} denotes the number of person-years at risk in the j th state of the kth cohort and $r_{jk}(x,v)$ denotes the corresponding mortality rate associated with a vector of covariates v, and a vector of potential confounders x.

2.1 Proportional Relative Risk Model

Under the proportional relative risk model, the mortality rate can be expressed as the product

$$r_{jk}\,(x,\, v) \;=\; r_{0jk}\,(x)\; RR_{jk}\,(v;\alpha_k), \tag{2}$$

where $r_{0jk}(x)$ denotes the background mortality rate for the j th state of kth cohort, $RR_{jk}(v;\alpha_k)$ is the associated relative risk, and α_k is a vector of model parameters.

In occupational cancer mortality studies, the association between cases and risk factors is often described by the linear exposure-response relationship

$$RR_{jk}(v;\alpha_k) \;=\; 1 \;+\; \beta_k \times w_j \times \xi_{k}, \tag{3}$$

where β_k is the slope parameter, w_j denotes the level of exposure to the risks factor of interest, and ξ_k is a vector of covariates which will modify the exposure-response relationship. When continuous exposure is stratified into L groups, the above relationship can be represented by categorical model

$$RR_{jk}\,(v;\alpha_k) = 1 + \beta_{Lk} \times \xi_k,$$

(l=1, ..., L), where β_{Lk} is the exposure-specific excess relative risk. Approximate methods for fitting non-linear models are discussed in sections 2.3 and 2.4; these methods are exact in the special case of linearity.

2.2 Fixed and Random Effects

Heterogeneity amount cohorts can be described by a random-effects model (Rutter and Elashoft, 1994), in which the overall effects and variation among individual cohorts are characterized by fixed and random regression coefficients respectively. Specifically, to describe heterogeneity across cohorts, the excess relative risk of lung cancer associated with exposure to radon for the kth cohort, β_k, is decomposed into two parts

$$\beta_k = \beta + b_{\beta,k}, \tag{4}$$

where β is the fixed effect for all cohorts combined and $b_{\beta,k}$ is the random effect specific to the kth cohort. In general, the parameters of model (3) or (4) can be characterized as

$$\alpha_k = \alpha + b_k \quad , \tag{5}$$

where $\boldsymbol{\alpha}$ denotes a vector of fixed effects applicable across all cohorts, and a vector of random effects $\boldsymbol{b}_k$ with zero mean specifies the deviation from the overall effects associated with the kth cohort.

2.3 Marginal Moments

More generally $RR_{jk}\,(v;\alpha_k)$ may be a non-linear function. To calculate the unconditional expectation and variance of the observed number of cases in random-effects model, we assume that the expectation of all random-effects is zero. Given b_k, the relative riskfunction can be approximated by

$$\begin{aligned} RR_{jk}(v;\,\alpha|b_k) &\approx RR_{jk}(v;\alpha) + \frac{\partial RR_{jk}(v;\alpha)}{\partial\alpha} \times b_k \\ &= RR_{jk}(v;\alpha)(1 + z_{jk}b_k)\ , \end{aligned} \tag{6}$$

where

$$z_{jk} = RR_{jk}^{-1}(v;\alpha) \times \frac{\partial RR_{jk}(v;\alpha)}{\partial \alpha}. \tag{7}$$

Given the random-effects b_k, the conditional expectation and variance of the observed number of deaths in the j th state of kth cohort for the Poisson regression model are

$$E(y_{jk}|b_k) = N_{jk}\, r_{jk}(x, v; \alpha|b_k) \tag{8}$$

and

$$Var(y_{jk}|b_k) = \phi E(y_{jk}|b_k). \tag{9}$$

where the over-dispersion parameter ϕ describes excess variation in the observed number of deaths. Letting $D = Cov(b_k)$ denote the covariance matrix of random effects, the marginal moments in the j th stratum of the kth cohort can be written as

$$E(y_{jk}) = \mu_{jk}(\alpha) - N_{jk}\, r_{0jk}(x)\, RR_{jk}(v;\alpha), \tag{10}$$

$$\begin{aligned} Var(y_{jk}) &= \sigma_{jk}(\alpha) - \phi\, \mu_{jk}(\alpha) \\ &\quad + \mu_{jk}^2(\alpha)\, z_{jk}^T(\alpha) D z_{jk}(\alpha). \end{aligned} \tag{11}$$

and

$$Cov(y_{jk}, y_{ik}) = \sigma_{jik}(\alpha) - \mu_{jk}(\alpha)\, \mu_{ik}(\alpha) z_{jk}^T(\alpha) D z_{ik}(\alpha) \tag{12}$$

$(j \neq i)$. The covariance matrix of random effects is an unknown non-negative definite matrix, which needs to be estimated in model fitting process. Let $Y_k = (y_1 ..., y_{J_k})^T$ be the vector of observations in the kth cohort, and let $\Omega = diag$ (**D**, ..., **D**) and $Z_k = diag(z_{1k}, ..., z_{J_k}k)$. Further, let $\Lambda_k(\alpha) = diag$ (μ_{1k} (α), ..., $\mu_{J_k k}$ (α)) and $\Sigma_k(\alpha) = \{\sigma_{ijk}(\alpha)\}$ be two $J_k \times J_k$ matrixes for which μ_{jk} (α) and σ_{ijk} (α) are given by (10) - (12). The covariance matrix for the vector of observations for the kth cohort (**Y** $_k$) can be written as

Table 1. Characteristics of 11 Underground Miners' Studies[a]

Location	Type of Mine	Number of Miners	Period of Follow-up	Number of Person-years	Number of Lung-cancer
China	Tin	17,143	1976-87	175,342	980
Czechoslovakia	Uranium	4,284	1952-90	107,868	661
Colorado	Uranium	3,347	1950-87	82,435	329
Ontario	Uranium	21,346	1955-86	380,718	291
Newfoundland	Fluorspar	2,088	1950-84	48,742	118
Sweden	Iron	1,294	1951-91	33,293	79
New Mexico	Uranium	3,469	1943-85	58,949	69
Beaverlodge	Uranium	8,486	1950-80	118,385	65
Port Radium	Uranium	2,103	1950-80	52,676	57
Radium Hill	Uranium	2,103	1948-87	51,850	54
France	Uranium	1,785	1948-86	44,043	45
Total		67,746	1943-91	1,151,315	2,736

[a]Cited from NCI report (Lubin *et al.*, 1994).

$$Cov(Y_k) = \Sigma_k(\alpha) - \phi\ \Lambda_k\ (\alpha)$$
$$+ \Lambda_k(\alpha)\ Z_k^T(\alpha)\,\Omega\, Z_k\ (\alpha)\ \Lambda_k(\alpha). \qquad (13)$$

2.4 Generalized Estimating Equations

Zeger *et al*., (1988) used a generalized estimating equations (GEE) approach to fit random-effects models to longitudinal data. The GEE approach permits relaxed distributional assumptions and is often computationally simpler than maximum likelihood estimation. With the specification of a working covariance matrix, the GEE method yields consistent and asymptotically normal estimates under mild regularity conditions, although efficiency may decrease.

Suppose that there are J_k states in the kth cohort (k=1,...,K). Letting $\mu_k(\alpha)$=E(Y$_k$), the GEEs for the fixed-effects α given the covariance of random-effects **D** = *Cov* (b$_k$) are

$$\sum_{k=1}^{K} \frac{\partial\mu_k^T(\alpha)}{\partial\alpha}\Sigma_k^{-1}\ (\alpha)(Y_k - \mu_k(\alpha))=0 \qquad (14)$$

(Zeger *et al*., 1988). This equation can be solved for α using Newton-Raphson iteration. Note that the estimating equation (14) is not unbiased since $\Sigma_k^{-1}(\alpha)$ is a function of the model parameters α (Burnett *et al*., 1995). For unbiased estimation of α, a penalty term needs to be added to (14) so that the expectation of the estimating equation is zero. The modified estimating equation has been referred as penalized quasi-likelihood function by (Breslow and Clayton, 1993). However, the inclusion of the penalty term adds to the computational burden of model fitting. In large samples, moreover, the bias may be negligible.

Following Zeger and Liang, (1988) the approximation (11) is used to obtain a preliminary estimate of the covariance matrix of random effects (D). Specifically, equation (11) can be re-expressed as

$$D - (z_{jk}z_{jk}^T)^{-1}z_{jk}(\frac{E(y_{jk}-\mu_{jk}(\alpha))^2 - \phi\mu_{jk}(\alpha)}{\mu_{jk}^2(\alpha)})z_{jk}^T(z_{jk}z_{jk}^T)^{-1}.$$

We use the moment estimator

$$\hat{D} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_k}\sum_{j=1}^{J_k}(z_{jk}z_{jk}^T)^{-1}$$
$$\{z_{jk}[\frac{(y_{jk}-\hat{\mu}_{jk}(\hat{\alpha}))^2 - \hat{\phi}\hat{\mu}_{jk}(\hat{\alpha})}{\hat{\mu}_{jk}^2(\hat{\alpha})}]z_{jk}^T\}\ (z_{jk}(z_{jk}^T)^{-1} \qquad (15)$$

to estimate the covariance matrix of the random-effects.

The moment estimator

$$\hat{\phi} = \sum_{k=1}^{K}\frac{1}{KJ_k}\sum_{j=1}^{J_k}\{\frac{(y_{jk}-\hat{\mu}_{jk}(\hat{\alpha})^2 - \hat{\mu}_{jk}(\hat{\alpha})z_{jk}\hat{D}z_{jk}^T}{\hat{\mu}_{jk}(\hat{\alpha})}\} \qquad (16)$$

is used to obtain an estimate of the over-dispersion parameter ϕ (Zeger and Liang, 1986). The estimates of α and (D, ϕ) are obtained by first solving (14) using Fisher scoring methods, with $(D,\ \phi)$ fixed at its estimated value $(\hat{D},\ \hat{\phi})$. Equations (14) are then evaluated with α fixed at its newly estimated value $\hat{\alpha}$, and iterating until convergence.

In this iterative process, the random effects related to the kth cohort can be estimated by using the estimate of $\hat{\alpha}$ to offset fixed effects in equation (5) of parameters of the relative risk model, and using the estimated covariance matrix of the random-effects $\hat{\Sigma}_k$ to solve the quasi-likelihood type score equation

$$\frac{\partial\mu_k^T(b_k)}{\partial b_k}\hat{\Sigma}_k^{-1}\ [Y_k - \hat{\mu}_k(b_k)] = 0 \qquad (17)$$

for $\mathbf{b}_k$ (k = 1, ..., K). Over-dispersion related to particular cohorts can be accommodated using methods suggested by McCullagh and Nelder (1989).

The estimates of the fixed-effects $\hat{\alpha}$, the estimates of random-effects $\hat{b}_k(k$=1,...,$K)$, and the covariance matrix of random-effects $\hat{D}$ are obtained by first solving (14) using Newton-Raphson methods, with D and $\mathbf{b}_k$ fixed at their initial values $\hat{D}$ and $\hat{b}_k$, respectively. Equations (15) and (16) are then evaluated with α fixed at its newly estimated value $\hat{\alpha}$, and iterating until convergence. Zeger and Liang (1986) note that convergent estimates (α^*) of the parameters are consistent, and suggest the robust variance estimate

$$Cov\,(\alpha^*) = \Gamma_0^{-1}\ \Gamma_1\ \Gamma_0^{-1}, \qquad (18)$$

where

$$\Gamma_0 = \sum_{k=1}^{K}\frac{\partial\hat{\mu}_k^T(\alpha^*)}{\partial\alpha^*}\hat{\Sigma}_k^{-1}(\alpha^*)\ \frac{\partial\hat{\mu}_k(\alpha^*)}{\partial\alpha^*}$$

and

$$\Gamma_1 = \sum_{k=1}^{K}\frac{\partial\hat{\mu}_k^T(\alpha^*)}{\partial\alpha^*}\hat{\Sigma}_k^{-1}(\alpha^*)\,[\,Y_k - \mu_k(\alpha^*)]^T$$

$$[\,Y_k - \mu_k(\alpha^*)]\,\hat{\Sigma}_k^{-1}(\alpha^*)\ \frac{\partial\hat{\mu}_k(\alpha^*)}{\partial\alpha^*}.$$

The variance estimates of $\hat{b}_k$ are

$$Cov(\hat{b}_k) = T_k^{-1} P_k T_k^{-1} \quad (19)$$

where

$$T_k = \frac{\partial \mu_k^T(\hat{b}_k)}{\partial \hat{b}_k} \hat{\Sigma}_k^{-1} \frac{\partial \mu_k(\hat{b}_k)}{\partial \hat{b}_k} \quad (20)$$

and

$$P_k = \frac{\partial \mu_k^T(\hat{b}_k)}{\partial \hat{b}_k} \hat{\Sigma}_k^{-1} [Y_k - \hat{\mu}_k(\hat{b}_k)]^T [Y_k - \hat{\mu}_k(\hat{b}_k)] \hat{\Sigma}_k^{-1} \frac{\partial \mu_k(\hat{b}_k)}{\partial \hat{b}_k}. \quad (21)$$

3. TWO-STAGE REGRESSION ANALYSIS

Depending on the number of categorical covariates, the number of states in the person-years table can be very large. This may present computational difficulties in model fitting. In this section, we discuss two-stage regression analysis methods which are easy to implement even though the number of confounders are large (Whitehead and Whitehead ,1991). Without loss of generality, we will use the simple linear model

$$RR_{jk}(\nu, \omega;\ \alpha_k) = 1 + \beta_k \times \omega_j \quad (22)$$

to illustrate the two step methods in which the parameters β_k denotes the excess relative risk for the kth cohort.

Stage 1. In the first stage, model (22) is fitted to each cohort. Let $\hat{\beta}_k$ be the estimate of model parameter β_k and s_k be the estimated variance of $\hat{\beta}_k$. The estimates $\{\hat{\beta}_k, s_k; k = 1, ..., K\}$ are used as input for Stage 2.

Stage 2. Define

$$\bar{\beta} = \frac{\Sigma_k s_k^{-1} \hat{\beta}_k}{\Sigma_k s_k^{-1}}, \quad (23)$$

$$\hat{\tau} = \frac{\Sigma_k s_k^{-1} (\hat{\beta}_k - \bar{\beta}_k)^2 - (K-1)}{\Sigma_k s_k^{-1} - \dfrac{\Sigma_k s_k^{-2}}{\Sigma_k s_k^{-1}}}, \quad (24)$$

and

$$w_k = \frac{(\hat{\tau} + s_k)^{-1}}{\Sigma_k (\hat{\tau} + s_k)^{-1}}. \quad (25)$$

Then the pooled estimate $\hat{\beta}$ of the overall effect for all cohorts is given by

$$\hat{\beta} = \sum_k w_k \hat{\beta}_k \quad (26)$$

The variance of the estimate of the overall effect is estimated by

$$Var(\hat{\beta}) = (\sum_k (\hat{\tau} + s_k)^{-1})^{-1}. \quad (27)$$

A statistical test for homogeneity of the $\hat{\beta}_k$ among cohorts is given by

$$\chi^2_{homog} = \sum_k s_k^{-1} (\hat{\beta}_k - \bar{\beta}_k)^2, \quad (28)$$

which has a chi-square distribution with K-1 degrees of freedom.

The shrinkage estimator of the cohort-specific effect β_k is given by

$$\hat{\beta}_k^* = \frac{s_k \hat{\beta} + \hat{\tau} \hat{\beta}_k}{s_k + \hat{\tau}}, \quad (29)$$

with the deviation from the overall estimate given by

$$\hat{\delta}_k = \hat{\beta} - \hat{\beta}_k^*, \quad (30)$$

provided $\hat{\tau} > 0$.

The estimated variance of the deviation is given by

$$Var(\hat{\delta}_k) = \frac{\hat{\tau} s_k}{\hat{\tau} + s_k}. \quad (31)$$

The heterogeneity among cohorts is described by τ: positive values of τ will increase the estimated variance of the overall effects $\hat{\beta}$.

4. ILLUSTRATION

In this section, the methods discussed in sections 2 and 3 will be used to analyze data from the 11 major studies of miners (Colorado, Czechoslovakia, China, Ontario, Newfoundland, Sweden, New Mexico, Beaverlodge, Port Radium, Radium Hill and France). A meta-analysis of these studies was conducted by Lubin *et al.*, (1994). We do not intent to conduct a comprehensive re-analysis the 11 underground miners' studies, but simply to illustrate the methods.

In total all these 11 studies include over 2,700 lung cancer cases among 68,000 miners representing nearly 1.2 million person-years of observation (Lubin *et al.*, 1994). Following Lubin *et al.*, the background lung cancer rate is stratified by age (all studies) defined by 5-year age groups (<40, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, ≥75 years), and other occupational exposures (available in the sudies from China, Ontario, Colorado, New Mexico and France), an indicator of radon progeny exposure (Beaverlodge), and ethnicity (New Mexico).

The relative risk model

$$E(y_{jk} \mid b_k) = N_{jk} r_{0jk}(x)[1+(\beta+b_k)\times w] \qquad (32)$$

Table 2. Estimated ERR/WLM %[a] Based on Random-Effects and Two-Stage Analyses

Study	Random Effects Analysis	Two Stage Analysis	Cohort Specific Analysis
Combined	0.49[b] ±(.14)[e]	0.56[c] ±0.12[e]	0.57[d] ±0.02[e]
China	0.25[f]	0.29[g]	0.28[h]
Czechoslovakia	0.75	0.67	0.71
Colorado	0.55	0.46	0.44
Ontario	0.52	0.73	0.94
Newfoundland	0.74	0.80	1.18
Sweden	0.53	0.62	1.68
New Mexico	0.64	0.60	2.66
Beaverlodge	0.53	0.64	4.26
Port Radium	0.42	0.42	0.38
Radium Hill	0.50	0.58	6.74
France	0.46	0.35	0.06

[a] ERR/WLM is the parameter β based on the model $RR = 1 + \beta \times w^*$, where $w^* = w_{5-14} + \theta_2 w_{15-24} + \theta_3 w_{25+}$ denotes cumulative radon exposure in which w_{5-14}, w_{15-24} and w_{25+} denote radon exposure received 5-14, 15-24, and more than 25 years ago, respectively. The values of θ_2 and θ_3 were estimated as 0.78 and 0.10, respectively.

[b] Fixed coefficient of random-effects model.

[c] Overall estimate of two-stage analysis.

[d] Weighted-mean of cohort-specific estimates. (Weights were inverses of the standarded errors of estimates.)

[e] The standardized error of estimate.

[f] Sum of fixed and random coefficients of random-effects model.

[g] Cohort-specific shrinkage estimate based on two stage analysis.

[h] Cohort-specific estimate.

will be used in this example, with the modifying effects of others covariates ignored for simplicity.

The parameter β_k reflects the excess relative risk per WLM radon exposure. Since radon exposures received more recently are considered more important than exposures received long ago the cumultative radon exposures in (22) is divided into three parts $w = w_{5-14} + \theta_2 w_{15-24} + \theta_3 w_{25+}$ in which w_{5-14}, w_{15-24} and w_{25+} denote radon exposure received 5-14, 15-24, and more than 25 years ago, respectively. The parameters θ_2 and θ_3 for the time-since-exposure effect are treated as fixed-effects with values of 0.78 and 0.10 respectively, as estimated from data from the 11 studies. The first column of Table 2 gives of estimates β_k based on all studies combined and the individual studies obtained using random-effects analysis techniques. The second column of Table 2 gives the corresponded estimates of β_k based on two-stage regression analysis methods. Last column of Table 2 gives the study-specific estimates which are based on conventional meta-analysis methods. The overall estimate of the excess relative risks based on the random-effects analysis is 0.49% with a standard error of 0.14%. The two-stage analysis leads to an estimate of 0.56% with a standard error of 0.115%. The conventional meta-analysis method provides an overall estimate of 0.57% with a standard error of 0.018%. The random-effects and two-stage analyses yield larger standard errors than the conventional analysis because heterogeneity between studies has been taken into account. Both the random-effects and two-stage analyses give similar estimates of the study-specific ERR shrunk towards the overall estimates of ERR.

5. CONCLUSION

In this article, we have presented methods for meta-analysis of a series of cohort studies used to investigate similar occupational health risks. Specifically both random-effects and two-stage analyses were discussed. A more conventional approach to meta-analysis based on a weighted linear combination of study specific estimates (with weights inversely proportional to the precision of the estimator) was also considered. Estimates of overall effects across all studies given by the three methods appeared comparable in the example used to illustrate the methods.

The random-effects analysis aggregates all available data, from all studies, whereas two-stage and conventional meta-analysis methods use only summary estimates from individual studies in synthesis. In contrast to conventional meta-analysis, heterogeneity between different studies is taken into account in estimating overall risks with both the random effects and two-stage analyses.

This advantage is particularly important when there are significant differences among the study results, that is, when the fixed-effects model fits the data poorly (Greenland, 1994). Computationally, the two-stage method can be more convenient than the random effects method with large datasets.

ACKNOWLEDGMENT

We would like to thank John D. Boice, Christer Edling, Richard W. Hornung, Geoffrey Howe, Emil Kunz, Robert A. Kusiak, Howard I. Morrison, Edward P. Radford, Jonathan M. Samet, Margot Trimarche, Alistair Woodward, Yao Shu Xiang, and Donald A. Pierce for permission to use the miners data in the example in section 4.

REFERENCES

Berlin, J.A., Longnecker, L.M., and Greenland, S. (1993). Meta-analysis of epidemiologic dose-response data, *Epidemiology*, 4, 218-228.

Berkey, C.S., Hoaglin, D.C., Mosteller, F., and Colditz, G.A. (1995). A random-effects regression model for meta-analysis, *Statistics in Medicine*, 14, 395-411.

Breslow, N.E., and Day, N.E. (1987) *Statistical Methods in Cancer Research,* Vol. 2: The Design and Analysis of Cohort Studies. International Agency for Research on Cancer, Lyon.

Burnett, R.T., Ross, W.H., and Krewski, D. (1995). Nonlinear mixed regression models, *Environmetrics,* 6, 85-99.

Greenland, S. (1994). Invited Commentary: A critical look at popular meta-analysis methods, *American Journal of Epidemiology*, 140, 290-296.

Greenland, S., and Longnecker, M.P. (1992). Methods for trend estimation from summarized dose-response data, with applications to meta-analysis, *American Journal of Epidemiology*, 135, 1301-1309.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, 38, 963-974.

Lubin, J.H., Boice, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Trimarche, M., Woodward, A., Xiang, Y.S., and Pierce, D.A. (1994). *Radon and Lung Cancer Risk: A Joint Analysis of 11 Underground Miners Studies*, National Institutes of Health, NIH Publication 94-3644.

Lubin, J.H. (1994). Invited commentary: lung cancer and exposure to residential radon, *American Journal of Epidemiology*, 140, 323-32.

McCullagh, P., and Nelder, J.A. (1989). Generalized linear models, Chapman & Hall, New York.

Moolgavkar, S.H. (1995). When and how to combine results from multiple epidemiological studies in risk assessment. Unpublished manuscripts (1992).

National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research*, National Academy Press, Washington, D.C.

Rutter, C.M., and Elashoff, R.M. (1994). Analysis of longitudinal data: random coefficient regression modelling, *Statistics in Medicine*, 13, 1211-1231.

Whittemore, A.S., McMillan, A. (1983). Lung cancer mortality among U.S. uranium miners: A reappraisal, *Journal of National Cancer Institute*, 71, 489-499.

Whitehead, A., and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials, *Statistics in Medicine*, 10, 1665-1677.

Zeger, S., Liang, K.Y., and Albert, P.S. (1988). Models for longitudinal data: A general estimating equation approach, *Biometrics*, 44, 1049-1060.

Zeger, S. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes *Biometrics*, 42, 121-130.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

LINKING DATA TO CREATE INFORMATION

W. Winkler and F. Scheuren[1]

ABSTRACT

The process of linking together records from two or more files has a long history, especially in Canada. Optimizing the linkage step and controlling the inevitable errors have been given primary consideration (Newcombe *et al.* 1959, Fellegi and Sunter 1969). Advances in estimating true linkage probabilities (*e.g.*, Belin and Rubin 1995, Winkler 1995) have been one of the engines needed for improved analysis. Our recent research (*e.g.*, Scheuren and Winkler 1993) on matched data in the presence of linkage errors has focussed on the interaction between linkage error and regression analysis – with adjustments being conditioned on what is known about the linkage. Here we continue that work but this time we are aiming at a more fully recursive approach. The first two steps would be the same as before: (1) estimating error probabilities and (2) adjusting the regression for linkage errors. The next step employs the analysis itself as an additional source of information. Outliers from the regression are treated as nonmatches with refitting of linkage probabilities and reestimation of the analysis relationships occurring as part of a recursive process.

KEY WORDS: Edit; Imputation; Record linkage; Regression analysis; Recursive processes.

1. INTRODUCTION

Researchers and policymakers often need more information than is available in a single data base. Use of microdata records from two or more data files, though, may be error prone when the only means of connecting corresponding records are nonunique identifiers, such as name and address information.

It is our view that recent developments in edit/imputation (**EI**) and record linkage (**RL**) have yielded tools of sufficient power that heretofore undoable analyses are now possible. Two of the reasons for this serve as the basis of the methods presented in this paper:

- First, while some problems exist, analysts generally can define the best ways to edit (*i.e.*, clean-up) and impute (*i.e.*, revise) existing microdata. We believe this is true with either individual data files or with merged files taken from several sources. What is new is that more individuals are applying powerful, reusable software routines based on edit models such as that of Fellegi and Holt (1976). Such models and software ease and systematize the **EI** process and replace **ad hoc**, data-base-specific **if-then-else** routines that must be written from scratch in each application.
- Second, analysts can effectively link a representative set of records from the separate data files via newly enhanced methods, originally introduced by Newcombe (Newcombe *et al.* 1959) and mathematically formalized by Fellegi and Sunter (1969). In earlier work (Scheuren and Winkler 1993), we demonstrated an **RL** adjustment procedure that reduced bias due to record linkage error in regression analyses.

In the present paper, we present a four-step recursive approach that is straightforward to carry out and very powerful. To start the process, we employ an enhanced **RL** approach (*e.g.*, Winkler 1995, Belin and Rubin 1995) to delineate a set of pairs of records in which the matching error rate is estimated to be very low. A regression analysis is attempted. Then, we use an **EI** model developed on the low-error-rate linked records to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (**RA**) is done and this time the results are then fed back into the linkage step so

[1] William Winkler, U.S. Bureau of the Census, Washington, DC 20225 and Fritz Scheuren, The George Washington University, Washington, DC 20056.

that the **RL** step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, we have

$$\begin{tikzcd} & RA \arrow[dr] & \\ RL \arrow[ur] & RA \arrow[l] & EI \arrow[l] \end{tikzcd}$$

Each of these three technologies, of course, is already in wide use. If this paper has something to contribute, it is as an illustration of one, we hope, sensible way to integrate them and thereby increase their utility even further.

Organizationally, the material is divided into five sections including this brief introduction. In the second section, we give a little background on edit/imputation and record linkage technologies. The empirical data files constructed and the regression analyses undertaken are described in Section 3. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

2. EDIT/IMPUTATION AND RECORD LINKAGE METHODS REVIEWED

In this section, we undertake a short review of Edit/Imputation (**EI**) and Record Linkage (**RL**) methods. Our purpose is not to describe them in detail but simply to set the stage for our present application. Because Regression Analysis (**RA**) is so well known, our treatment of it is covered only in the particular application (Section 3).

2.1. Edit/Imputation

Historically, methods of **editing** microdata arose mainly to deal with logical inconsistencies in data bases (*e.g.*, Nordbotten, 1963); their use in detecting unlikely or implausible entries has also been important (*e.g.*, Granquist, 1984). Methods of **imputing** microdata had their beginnings mainly as a way of handling missing entries (*e.g.*, Little and Rubin, 1987). Using the two in combination has long been done, though, by practitioners. Edit/ Imputation (**EI**) methods were, however, not fully conceptualized as a single system until the seminal paper by Fellegi and Holt (1976). Attempts at completely computerizing the Fellegi-Holt system had to wait even longer. It is only now with the present generation of computer hardware and software that success can be said to be satisfactory.

Although we will only consider continuous data in this paper, **EI** techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case a collection of edits might consist of rules for each record of the form

$$c_1X < Y < c_2X \qquad (2.1)$$

In words,

If Y less than c_1X or greater than c_2X, then the data record should be reviewed. (2.2)

Here **Y** may be total wages, **X** the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$.

Methods of editing have consisted of sets of **if-then-else** rules, statistical methods based on outlier tests detection, and the formal model of Fellegi and Holt (1976) which generally contains other types of models as special cases. The main advantages of the Fellegi-Holt model are that: (1) it allows the systematic checking of a system of edits for logical consistency prior to the receipt of data, (2) it determines the minimum amount of information that must be changed in a record so that the revised record satisfies all edits, and (3) it is table-driven with edit restraints residing in tables and the main edit routines being reusable. For items that need to be changed, an imputation algorithm is integrated into the system that satisfies the edit checks and which can be used to replace entries found inconsistent.

The current general Fellegi-Holt systems that run on a variety of computers consist of Statistics Canada's GEIS (Generalized Edit and Imputation System) for linear inequality edits (*e.g.*, Kovar, Whitridge, and MacMillan 1991); the Census Bureau's new SPEER (Structured Programs for Economic Editing and Referral) system for ratio edits of continuous data (*e.g.*, Winkler and Draper 1996), and the Census Bureau's DISCRETE system for edits of general discrete data (*e.g.*, Winkler and Petkunas 1996). Imputation is via now standard techniques (*e.g.*, Little and Rubin 1987) -- often a variant of the "Hot Deck."

2.2. Record Linkage

A record linkage process attempts to classify pairs in a product space **A × B** from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.*, 1959), Fellegi and Sunter (1969) considered ratios **R** of probabilities of the form

$$\mathbf{R} = \Pr(\gamma \in \Gamma \mid \mathbf{M}) / \Pr(\gamma \in \Gamma \mid \mathbf{U}) \qquad (2.3)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called matching variables.

The decision rule is given by:

If R > *Upper*, then designate pair as a link.
If *Lower* ≤ R ≤ *Upper*, then designate pair as a possible link and hold for clerical review. (2.4)
If R < *Lower*, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that for any pair of fixed bounds on **R**, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio **R** or any monotonely increasing transformation of it (typically a logarithm) a matching weight or total agreement weight.

Like **EI** methods, **RL** techniques have made major advances as an offshoot of cheap, available computing. Over about the last decade, there has been an outpouring of new work on record linkage techniques (*e.g.*, Jaro, 1989; Newcombe, Fair, and Lalonde, 1992). Some of these results were spurred on by a series of conferences beginning in the mid 1980s (*e.g.*, Kilss and Alvey, 1985; Carpenter and Fair, 1989); a further major stimulus in the U.S. has been the effort to study undercoverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau, 1991). The seminal book by Newcombe (1988) has also had an important role in this ferment.

3. SIMULATION SETTING

For our simulations, we considered four matching scenarios as in our earlier work (Scheuren and Winkler 1993). The basic idea was to generate data having known distributional properties, adjoin the data to two files that would be matched, and then to evaluate the effect of increasing amounts of matching error on analyses. We started with two files (of size 12,000 and 15,000) having good matching information and for which true match status was known. About 10,000 of these were true matches (before introducing errors) -- for a rate on the smaller or base file of about 83%.

We then generated empirical data with known distributional properties and adjoined the data to the files. As we conducted the simulations, a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed. These variations are described below and shown in figure 1. For each scenario in the figure, the match weight, the logarithm of **R**, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o):

Good Scenario (figure 1a). -- Previously, we had concluded that no adjustments for matching error are necessary here. This scenario can happen in systems designed for matching, having good matching variables, and that use advanced matching algorithms. The true mismatch rate here was under 2%.

Mediocre Scenario (figure 1b). --The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but considered consistent with those that might be selected by an experienced computer matching expert. The true mismatch rate here was 6.8%.

First Poor Scenario (figure 1c). -- The first poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience. The true mismatch rate here was 10.1%.

Second poor Scenario (figure 1d). --The second poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names. Severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. The true mismatch rate was 14.6%.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 1a). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 1b); with the first poor scenario, the overlap is substantial (Figure 1c); and, with the second poor scheme, the overlap is almost total (Figure 1d).

RL true mismatch error rates can be reasonably well estimated by the procedure of Belin and Rubin (1995), except in the second poor scenario where the Belin-Rubin procedure would not converge. In practice, for this scenario there is almost no part of the data for which true link status would be known without followup operations. Until now an analysis based on the second scenario would not have been seemed even remotely sensible. As we will see in Section 4, something of value can be done, even in this case.

Having specified the above linkage situations, we then used SAS to generate ordinary least squares data under the model $\mathbf{Y} = \mathbf{4}\,\mathbf{X} + \epsilon$. The **X** values were chosen to be uniformly distributed between 1 and 101 and the error terms ϵ are normal and homoscedastic with variance 4000 -- all such that the regression of **Y** on **X** has an R^2 value in the true matched population of 78%. Only the results for the second poor scenario are presented here in detail. It's results were far and away the most dramatic.

4. RECURSIVE PROCESSES AND RESULTS

Graphs of the recursive process for the second poor scenario are discussed here. The regression results are given for two cycles. Furthermore, to help sort out what is happening the plots are displayed in separate panels at each step.

4.1 First Cycle Results

4.1.1 True regression (for reference). -- Figure 2 is a scatterplot of **X** and **Y** as they would appear if there were no matching errors. Note all of the mismatches are plotted but only 5% of the true matches are being used. This has been done to keep the true matches from dominating the results so much that no movement can be seen. Second, in this figure and throughout the remaining ones, the true regression line is always given for reference. Finally, the true population slope or **beta** coefficient (at 3.99) and the R^2 value (at 78%) are provided for the data being displayed.

4.1.2 Regression after Initial **RL** ⇒**RA** Step. -- In figure 3, we are looking at the regression on the actual observed links -- not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between **Y** and **X**. The observed slope or **beta** coefficient differs greatly from its true value (1.18 v. 3.99). The fit measure is similarly affected, falling to 7% from 78%.

4.1.3 Regression after Combined **RL**⇒**RA**⇒**EI**⇒**RA** Step. -- Figure 4 completes our display of the first cycle of our recursive process. Here we have edited the data in the plot displayed as follows. First, using just the 183 cases with a match weight of 3.00+, an attempt was made to improve the poor results given in figure 3. Using this provisional fit, predicted values were obtained for all the matched cases; then outliers with residuals of 100 or more were removed and the regression refit on the remaining pairs. This new equation was essentially $\mathbf{Y} = 3\mathbf{X} + \epsilon$ with a standard deviation of 3000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the **beta** coefficient from 3.0 to 3.4. If a pair of matched records yielded an outlier, then predicted values using the equation $\mathbf{Y} = 3.4\mathbf{X}$ were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

4.2 Second Cycle Results

4.2.1 True regression (for reference). -- Figure 5 displays a scatterplot of **X** and **Y**, as they would appear if they could be true matches based on a second **RL** step. The second **RL** step employed the predicted **Y** values as determined above; hence it had more information on which to base a linkage. This meant that a different group of linked records was available after the second **RL** step. In particular, since a considerably better link was obtained, there were fewer false matches; hence our sample of all false matches and 5% of the true matches dropped from 1606 in figures 2 thru 4 to 1104 for figures 5 thru 7. In this second iteration, the true slope or **beta** coefficient and the R^2 values remained, though, virtually identical for the slope (3.94 v. 3.99) and fit (77% v. 78%).

4.2.2 Regression after second **RL** ⇒ **RA** Step. -- In figure 6, we see a considerable improvement in the relationship between **Y** and **X** using the actual observed links after the second **RL** step. The slope has risen from

.18 initially to 3.64 here. Still too small but much improved. The fit has been similarly affected, rising from 7% to 65%.

4.2.3 Regression after Combined RL→RA→EI →RA Step. -- Figure 7 completes the display of the second cycle of our recursive process. Here we have edited the data as follows. First, using just the 185 cases with a match weight of 7.00+, an attempt was made to further improve on the results obtained in figure 6. Using this fit, another set of predicted values was obtained for all the matched cases. This new equation was essentially **Y** = 3.8**X** + ϵ with a standard deviation of about 2000. Using our earlier approach (Scheuren and Winkler 1993), a further adjustment was made in the **beta** coefficient from 3.8 to 4.0. Again, if a pair of matched records yields an outlier, then predicted values using the equation **Y** = 4.0**X** were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value. The plot in figure 7 gives the adjusted values.

5. CONCLUSIONS AND AREAS FOR FUTURE STUDY

In principle, the recursive process outlined above could have continued. Indeed, in a real problem, we would have had to continue until the **Y** v. **X** relationship (i.e., the **beta** coefficient in this example) ceased to change appreciably.

At first it would seem that we should be happy with the results. They take a seemingly hopeless situation and give us a fairly sensible answer. Would this always happen? We think so but this is an area we will be looking at in our future work. A closer examination, though, shows a number of places where the approach taken is weaker than it needs to be or simply unfinished.

5.1 Overfitting

The algorithm we are using has overfit the relationship between **X** and **Y**. The R^2 value after the second cycle was 84% versus 77% in the population. This is a common problem in regression when outliers are removed and we might be willing to be simply philosophical about it. However, there are several things that could be done at the **EI** step to improve our procedure.

The one we are most attracted to is to employ an idea of Howard Newcombe's and use a sample of **known** nonmatches. In our earlier work (Scheuren and Winkler 1993) on this problem, the matching we proposed involved getting two links for each case in the base file. The second link would be a next best and usually could be assumed to be a false match. How could these second matches be used to help with the overfitting problem?

Assume first that the matching is good enough so that the Belin-Rubin algorithms work (Belin and Rubin 1995) and we can calculate a true link probability for each match. We would then be able to estimate the number of false links among our best matched cases. This number of cases could be selected from the second best match file -- perhaps simply at random or better in a balanced way (such that, say, the means of **X** and **Y** in this false matched sample file agreed with the corresponding values in the original or best match file). A possible next step here would be to match the false matched sample to the original best matched cases and remove the "closest" pairs. This would be done instead of looking for outliers and removing all those at some distance from the center of the data (as was described in 4.1.3).

Even if the Belin-Rubin algorithms do not converge on the first cycle, Newcombe's idea of using a file of nonmatches might still be tried once the recursive process yielded matches of sufficient quality to employ it. In the present example this would have been possible at the second cycle, even though it was not possible initially.

5.2 Diagnostics

Under the assumptions, so far implicit, in our simulations, we are treating the matching variables and their relationship from file to file as independent of the **(X, Y)** relationship. A fuller discussion of this has been given in our earlier work (Scheuren and Winkler 1993). It is enough here to indicate that sample diagnostics should be tracked to check on this assumption. At each stage, it makes sense to calculate certain univariate statistics from the cases being treated as matched at that stage -- means and medians for **X** and **Y**, even the mean squares and mean deviations for **X** and **Y**. Using these quantities to protect against overfitting might even be possible. This will be an area for future study.

5.3 A Sampling of Open Issues

In the detailed working out of our approach, several **ad hoc** fixes were built into the **EI** step. How would one decide, for example, where to cutoff the matched cases so as to get a provisional regression? What about using a median trace as the starting point against which to identify outliers? Why was the outlier cutoff set where it was? Would a looser bound have helped appreciably in reducing the overfitting?

Figure 1a. Good Matching Scenario

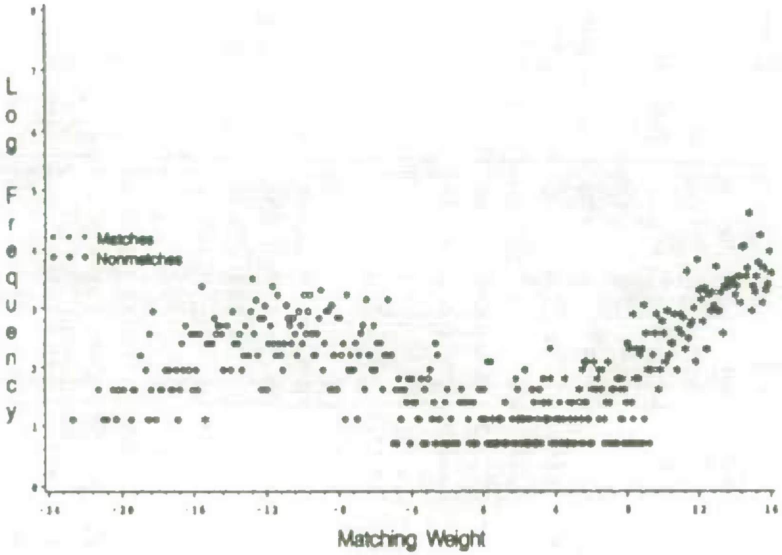


Figure 1b. Mediocre Matching Scenario

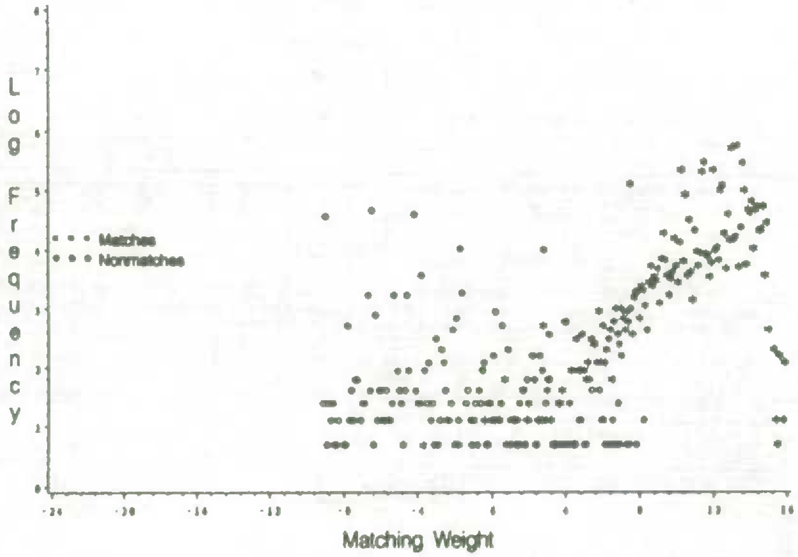


Figure 1c. 1st Poor Matching Scenario

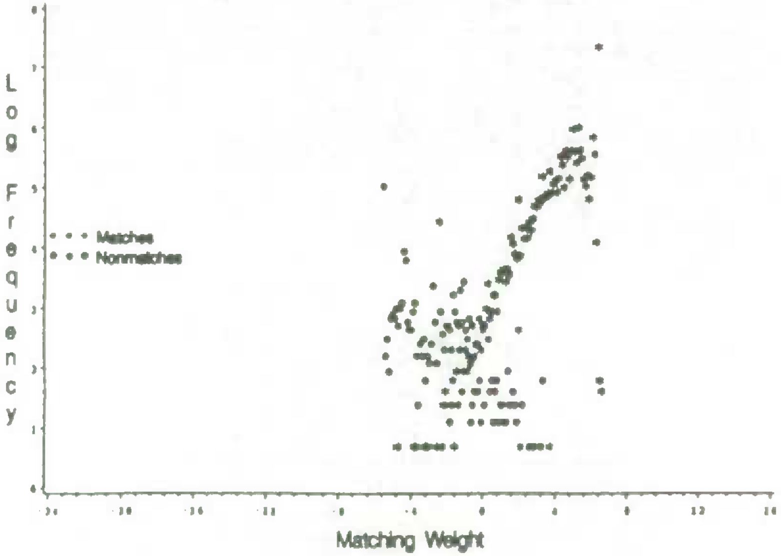


Figure 1d. 2nd Poor Matching Scenario

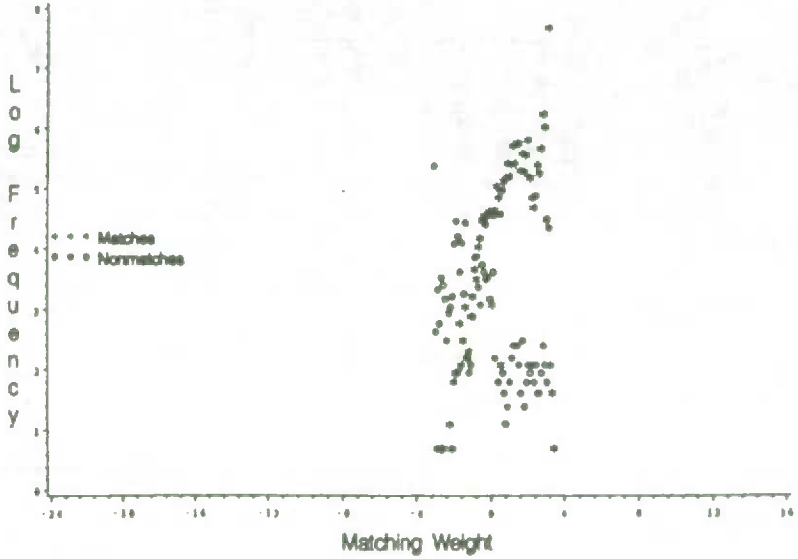


Figure 2. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, True Regression Data
1606 data points, beta = 3.99, R - square = 0.78

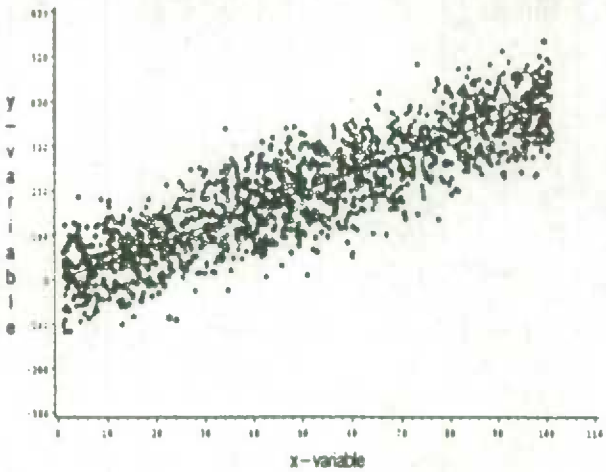


Figure 3. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, Observed Data
1606 data points, beta = 1.18, R - square = 0.07

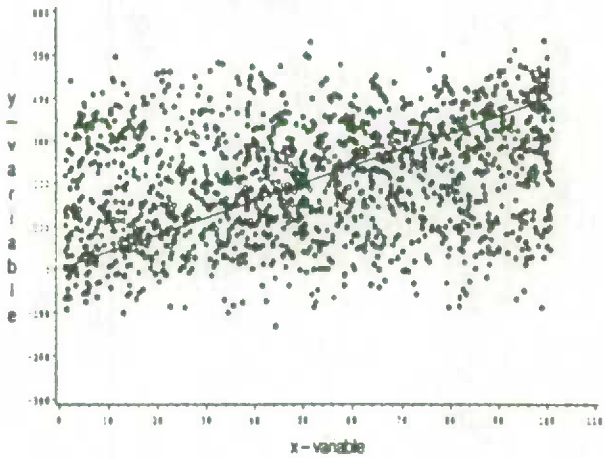


Figure 4. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, Outlier - Adjusted Data
1606 data points, beta = 3.46, R - square = 0.75

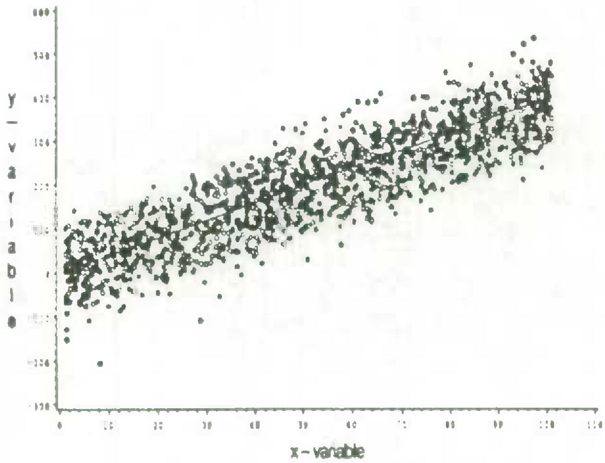


Figure 5. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, True Regression Data
1104 data points, beta = 3.94, R - square = 0.77

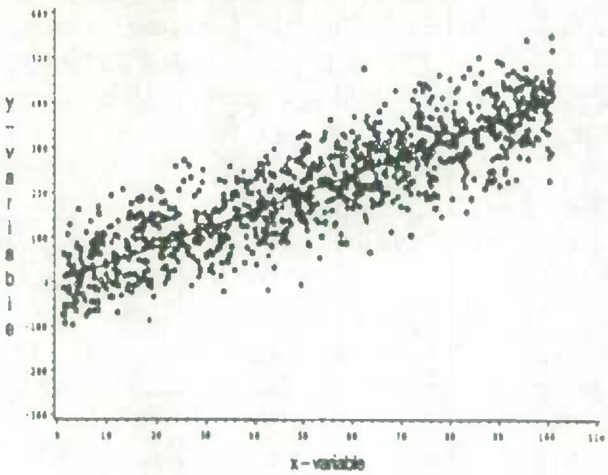


Figure 6. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, Observed Data
1104 data points, beta = 3.64, R - square = 0.65

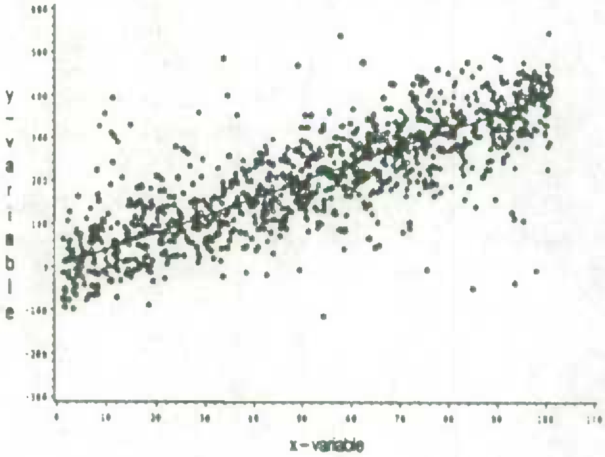
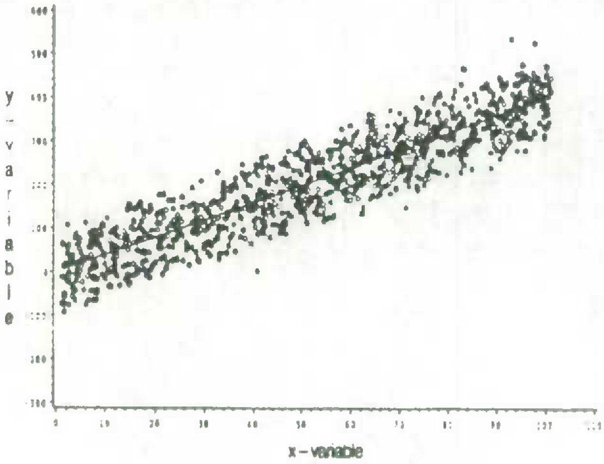


Figure 7. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, Outlier - Adjusted Data
1104 data points, beta = 4.01, R - square = 0.84



There were also **ad hoc** elements in the **RL** step. Already discussed was trying to introduce the Belin-Rubin algorithm as soon as possible. A lot more reflection is needed on the way the **RA** results were used, too, from the first cycle in the second **RL** step. For example, why not use the fitted regression value in every case not just for the outliers?

5.4 Generalizability Concerns

We have looked at a simple regression of one variable from one file with another variable from another. What happens when this is generalized to the multiple regression case? We are working on this now and feel sensible results will emerge but stay tuned.

On some other issues we are less sure about how to generalize. For example, what happens when the relationship between **Y** and **X** is weak in the population. Maybe then we cannot improve the match enough to make all the work being done here worthwhile? What happens when the overlap between the two files is very low (it was high in our example)?

5.5 Statistical Technology and Statistical Theory

This paper has been about technological possibilities. Our discussion has not been independent of theoretical considerations; but, conversely, the theoretical underpinnings of the ideas being explored have not all been worked out either. This early, intuitive approach is not unexpected of work in progress. We do not apologize for it; rather, in some ways it allows you, the listener (or reader), to become a player. One of our goals in our earlier work was to get others involved. It is our goal again.

5.6 From Data to Information

At the presentation, there was a lot more discussion than has been given here of the need for a recursive process that united housekeeping efforts, like getting good links, with analysis ones, like fitting a regression relationship. We feel strongly that now that the computer power and related software are becoming available the data producer and data user roles both need to change. Each needs to become much more interactive -- a real team effort. There is a price to pay for this change but there is an even bigger price to pay if we continue to be as separate as we are.

REFERENCES

Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage, *Journal of the American Statistical Association*, 90, 694-707.

Carpenter, and Fair, M.. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, in Ottawa, Ontario, Canada, August 30-31, 1989, Statistics Canada.

Fellegi, I., and Holt, T.(1976). A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association*, 71, 17-35.

Fellegi, I., and Sunter, A. (1969). A theory of record linkage, *Journal of the American Statistical Association*, 64, 1183-121

Granquist, L. (1984). On the role of editing, *Statistic Tidshrift*, 2, 105-118.

Jabine, T. B., and Scheuren, F. J. (1986). Record linkages for statistical purposes: methodological issues, *Journal of Official Statistics*, 2, 255-277.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.

Kilss, B., and Alvey, W., (Editors) (1985). *Record linkage techniques- 1985*, U.S. Internal Revenue Service, Publication 1299 (2-86).

Kovar, J.G., Whitridge, P., and MacMillan, J. (1988). Generalized edit and imputation system for economic surveys at Statistics Canada, *American Statistical Association, Proceedings of the Section of Survey Research Methods*, 627-630.

Little, R.J.A., and Rubin, D.B., (1987). *Statistical Analysis with Missing Data*, New York: John Wiley.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). Automatic linkage of vital records, *Science*, 130, 954-959.

Newcombe, H., Fair, M., and Lalonde, P. (1992). The use of names for linking personal records, *Journal of the American Statistical Association*, 87 1193-1208.

Neter, J., Maynes, E.S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors, *Journal of the American Statistical Association*, 60, 1005-1027.

Nordbotten, S. (1963). Automatic editing of individual observations, presented at the Conference of European Statisticians, UN Statistical and Economic Commission of Europe.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched, *Survey Methodology*, 19, 39-58.

Tepping, B. (1968). A model for optimum linkage of records, *Journal of the American Statistical Association*, 63, 1321-1332.

Winkler, W.E. (1995). Matching and record linkage, in B.G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W.E., and Draper, L. (1996). Application of the SPEER edit system, in *Statistical Data Editing, Volume 2*, Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

Winkler, W.E., and Petkunas, T. (1996). The DISCRETE edit system, in *Data Editing, Volume 2*, Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

Winkler, W., and Thibaudeau, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census, *Statistical Research Division Technical Report*, U.S. Bureau of the Census.

SESSION 2

Analytical Methods

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

SOCIO-ECONOMIC STATISTICS AND PUBLIC POLICY: A NEW ROLE FOR MICROSIMULATION MODELING

M.C. Wolfson[1]

ABSTRACT

Users of socio-economic statistics typically want more and better information. Often, these needs can be met simply by more extensive data collections, subject to usual concerns over financial costs and survey respondent burdens. Users, particularly for public policy purposes, have also expressed a continuing, and as yet unfilled, demand for an integrated and coherent system of socio-economic statistics. In this case, additional data will not be sufficient; the more important constraint is the absence of an agreed conceptual approach.

In this paper, we briefly review the state of frameworks for social and economic statistics, including the kinds of socio-economic indicators users may want. These indicators are motivated first in general terms from basic principles and intuitive concepts, leaving aside for the moment the practicalities of their construction. We then show how a coherent structure of such indicators might be assembled.

A key implication is that this structure requires a coordinated network of surveys and data collection processes, and higher data quality standards. This in turn implies a breaking down of the "stovepipe" systems that typify much of the survey work in national statistical agencies (i.e. parallel but generally unrelated data "production lines"). Moreover, the data flowing from the network of surveys must be integrated. Since the data of interest are dynamic, the proposed method goes beyond statistical matching to microsimulation modeling. Finally, these ideas are illustrated with preliminary results from the LifePaths model currently under development in Statistics Canada.

KEY WORDS: Microsimulation; Social statistics; Statistical frameworks.

1. INTRODUCTION

It is an eminently reasonable expectation that a nation's statistical system provide reliable, coherent, and salient views of central socio-economic processes (e.g. Garonna, 1994; OECD, 1976). To an important extent, this is accomplished by the System of National Accounts (SNA). However, it has also long been appreciated that the SNA suffers from many serious limitations, particularly from the viewpoint of social concerns and policy. These limitations are implicit in the history of attempts to create sets or systems of internationally agreed social indicators.

To date, nothing has emerged from efforts to construct social indicators that compares to the SNA in breadth, coherence, and international acceptance. As a result, fundamentally new approaches appear necessary to meet these needs for socio-economic statistics, including the needs that have long motivated efforts in the area of social indicators.

Broadly speaking, three main strategies have been proposed for developing a statistical framework for the

[1] Michael C. Wolfson, Statistics Canada, Institutions and Social Statistics, 24th floor, R.H. Coats Bldg., Ottawa, Ontario, K1A 0T6, Canada, e-mail:wolfson@statcan.ca.
Statistics Canada and the Canadian Institute for Advanced Research. The work reported here is very much a team effort principally by members of the Socio-Economic Modeling Group of the Analytical Studies Branch, Statistics Canada. Geoff Rowe and John Armstrong led the empirical analysis, while Steve Gribble created the ModGen software environment within which the LifePaths microsimulation model is written, and provided continuing leadership to the team. I remain responsible for any errors or infelicities in this paper.
A paper prepared for the 50th Session of the International Statistical Institute, Beijing, 1995.
This paper represents the views of the author and does not necessarily reflect the opinions of Statistics Canada.

social sphere. One is extensions of the SNA, most prominently in either the form of Social Accounting Matrices (SAMs, e.g. Pyatt, 1990), or Satellite Accounts (e.g. Vanoli, 1994; Pommier, 1981). The second proposed strategy is construction of a framework designed specifically for social statistics -- the best known and most clearly articulated being Stone's System of Social and Demographic Statistics (SSDS; UN, 1975; Stone, 1973). The third approach foregoes the structure and coherence of an explicit framework, seeking only consensus on an ad hoc collection of statistical indicators. This is exemplified by the set of social indicators recommended by the OECD (Moser, 1973; OECD, 1976).

All three strategies have failed to achieve broad implementation within advanced countries, and have as a result failed to provide a basis for internationally comparable data. Three reasons can be identified for these failures. One is concern about feasibility, a second is lack of priority from national governments or statistical agencies; and a third is lack of salience . (Obviously, these reasons may be related.) The Social Indicators experience (OECD, 1982) did result in a completely specified and operational list of indicators -- based on agreement among both experts and senior government representatives of member countries. However, for many of the agreed indicators (e.g. healthfulness of life, use of time, income), the requisite data collection systems have still not been created in an internationally comparable manner. Since data systems for these purposes exist within some countries, it is clearly not a question of technical feasibility.

This leaves some mix of lack of salience and lack of sufficient priority -- including unwillingness to invest the necessary resources in statistical data collection -- as the main explanations. To the extent there is a lack of salience for social indicators, it is more likely in the area of international comparability than in domestic usefulness. Most countries do have a wide range of social statistics; the problem is that they are generally not comparable from one country to another. This revealed lack of interest in a comprehensive set of internationally comparable social statistics may simply be a result of the fact that economic concerns (as reflected in the successful efforts to create an internationally agreed SNA) are broadly seen as much more important than social concerns, at least from a comparative point of view. Another might be that a handful of basic social indicators (e.g. life expectancy, unemployment rates) are already available on an internationally comparable basis, and these are felt sufficient.

One source of relevance of internationally comparable data is that countries feel strong connections with one another in the given domain. In the economic sphere, it is obvious that countries are tied together substantively by trade and international financial flows, and by shared intellectual roots in macro-economic theory. These connections form a basis for the SNA. Corresponding substantive connections in the social sphere may simply be seen to be weaker (though cultural and intellectual flows via mass media and coincident international patterns of terrorism, unemployment, marriage breakdown, fertility decline, growing wage inequality, etc. make this view tenuous). In addition, if there is no shared theory, the basis for international comparability is weaker -- both at the level of individual statistical series (e.g. household income distribution), and how various social statistics series are put together (if at all). In short, countries' failures to invest in creating the data collection systems necessary for the OECD's Social Indicators may be the result of a lack of interest in comparability among an ad hoc collection of indicators.

However, the failures of implementation of the SSDS or Satellite Accounts in the social area must be due to more than a lack of interest in internationally comparable data; they must also derive from a lack of interest in their implicit underlying theoretical frameworks. The failures here are more profound because even within countries, it is difficult to find coherent and comprehensive structures for social statistics that come anywhere close to the SNA in scope and implementation.

The decades of failure of all of these strategies suggests that new approaches are required in order to provide a coherent structure for social statistics. Such coherence would help meet the needs of a growing segment of users. On one side, it would provide the basis for generalist needs, for example summary indications of broad trends. On the other side, bringing coherence to the complex of social statistics should help more specialized users who can currently become confused (justifiably) when they find several inconsistent estimates for the same item from different data sources.

2. STARTING FROM FIRST PRINCIPLES

As a first step in thinking about new approaches to the construction of a framework for social statistics, it is useful to set out a series of basic measurement objectives. There are three which hopefully command broad support.

a. **overall outcome** -- A primary objective is to tell whether or not "things are getting better all the time" (in the words of the Beatles). Are people better off than they were last year, than a decade ago? Answering this question is difficult, primarily because there is no widely agreed summary approach to measuring how well off individuals are. Money income, health status, educational attainment, and social deprivation are all ingredients for such a measurement. But there is no agreement as to what other ingredients are essential, nor how the ingredients should be combined into an overall index. Moreover, there is lack of agreement on the appropriate outcome concept within various domains such as health and education.

The character of this partial consensus has major implications for statistical framework development. One is that flexibility is required. To the extent that there is agreement on some of the ingredients in measuring overall well-being, these ingredients must be included in the underlying statistical program. But given that there is no single "right answer" as to how they should be combined, users should have flexibility in combining them. It should be possible to assemble alternative summary indices from the basic ingredients -- both at the level of domains like health and education, and overall.

A second implication is that social statistics should be neither subservient to nor separate from economic statistics, as has characterized the three main strategic approaches so far. Economic status is clearly a central ingredient in any broad measure of whether or not people are becoming better off. Thus, it is better to think of a statistical framework for socio-economic statistics than for social statistics alone. In turn, and in contrast to the views of some national accountants (e.g. Vanoli, 1994), the prospects for frameworks for social statistics that build out from SNA concepts are not good. Completely fresh approaches are required. They should build on new premises, taking the whole of the "social economy" as the domain. The SNA would then become an important component of a broader "System of Social and Economic Statistics" (Wolfson, 1994; Ruggles and Ruggles, 1973).

Yet a third implication is that summary methods for comparing two or more social economies -- either over time or between countries -- must be found that go beyond simple linear aggregation. The concepts of interest are not always amenable to aggregation based on a single numeraire as is the case with money values in the SNA. Fortunately, work in areas like methods for comparing household income distributions, or distributions more generally using graphical methods (e.g. Easton and McCulloch, 1990), and flexible data base architectures allowing set-theoretic and pointer-based arithmetic over complex multivariate, longitudinal micro data sets -- which are now readily feasible with modern informatics -- indicate that SNA-style aggregation is not essential. Indeed, such approaches complement and support the second following objective.

b. **variety**-- Another basic measurement objective is to enable users to see the variety in the social economy. Variety covers all kinds of heterogeneity -- for example going beyond aggregates and -averages to meet the long-standing criticism of the SNA that it conveys nothing of the rich, the poor, and the degree of inequality in the size distribution of income. Variety is also reflected in the dispersion of educational attainments and household structures in a country's population.

Capturing variety has fundamental implications for a statistical system. Essentially, it requires explicit micro data foundations. The SNA, as the pre-eminent statistical framework, predates the revolution in computing. This has evidently constrained creative thinking for social and socio-economic statistical frameworks. In the pre-computer era in which the basic structure of the SNA was developed, aggregation was not only part of the theoretical foundation, it was also a practical necessity. Today with modern database technology, aggregation is not only inimical to accurate reflection of variety, it is also practically unnecessary.

(The idea of explicit micro data foundations in such statistical endeavors is certainly not new, e.g. United Nations, 1979; Ruggles, 1981. On the other hand, the pervasiveness of the "aggregation culture" is evident in the OECD Social Indicators efforts wherein it was felt necessary to define a set of a priori "basic disaggregations of main social indicators"; OECD, 1977. Such agreement, while helpful, is certainly not essential when appropriate and internationally comparable micro data sets are readily available to analysts.)

c. **what if** -- The third fundamental measurement objective is to provide the basis for the careful posing of and answering "what if" questions. There are two basic motivations. The obvious one is that government policy departments and private sector decision-makers, as major users of socio-economic statistics, want answers to these kinds of questions. For example, what would the distribution of disposable incomes be if such and such a tax/transfer policy change were implemented; or what would spending on such and such commodities be in five years if current trends prevailed?

Less obvious but equally important is that key statistical indicators are in fact answers to such "what if" questions. The best example is life expectancy. Life expectancy in 1990, say, is the answer to the following hypothetical question, "how long could a birth cohort expect to live if all its members were always exposed to the mortality rates that were observed in 1990?" Thus, life expectancy is not a datum that is directly observed like counts of deaths by age and sex. Rather, it is the outcome of a numerical simulation, one that is tightly coupled to disaggregated numerator and denominator count data on deaths and populations at risk respectively.

While life expectancy is a hypothetical construct that is not directly measured, it is also an intuitive and broadly accessible concept that can provide the framework for related families of indicators. This is clearest in the health area, where an ad hoc group of researchers has come together in the REVES (reseau pour esperance de vie en sante; Mathers and Robine, 1993) to develop and seek consensus on just such a related family of health indicators.

Drawing this discussion together, three basic measurement objectives have been proposed:

- broad indicators of the extent to which individuals are becoming better off;
- capacity to display variety and heterogeneity in the population; and
- tools for posing and answering "what if" questions.

In turn, in order to meet these objectives, the underlying statistical framework must:

- be flexible;
- encompass both social and economic aspects;
- have explicit micro data foundations;
- utilize modern informatics and database technology; and
- incorporate simulation models that are tightly coupled to data.

Given these premises, what might serve as the key elements of a framework for socio-economic statistics? The following points are apt:

- at any point in time, the population is best represented by a sample of individuals, each of whom is characterized by a set of attributes and relationships;
- attributes include income, educational attainment, consumption, various aspects of health status, and time use patterns of activity;
- relationships include conventional kinship ties as well as cohabitation (i.e. in database or graph-theoretic terms, such relationships can be represented by various kinds of pointers to other individuals -- each of whom is also in the database);
- relationships also include interactions with the major institutions of society -- school, work and government programs. These contacts, relationships or transactions between individuals and major institutions can also be considered part of the set of individual attributes. They can take the form of pointers to descriptions of the institutions -- schools, workplaces, and government programs -- with which the individuals were interacting;
- this individual database can then easily be viewed as comprised of a hierarchy of various types of units -- e.g. individuals, nuclear families, extended families, and households;
- each unit (individual, family or household) can be described by any one of a number of summary attributes such as disposable income, leisure time, or self-reported satisfaction;
- measures of variety can then be defined by summary statistics over this multivariate joint distribution of units (e.g. Gini coefficients, quantile shares);
- over time, the population is best represented by a series of individual biographies, the equivalent of a broad and deep longitudinal panel survey;
- given this longitudinal representation, a coherent family of summary indicators can be constructed from generalizations of the notion of life expectancy -- including partitions of life expectancy into cumulative sojourn times in various life states;

In essence, this socio-economic framework would contain a complete longitudinal micro data sample, a microcosm of the actual population and its relationships to major social and economic institutions. From this microcosm, a wide variety of statistical indicators could be readily constructed -- effectively with no more effort than pressing the ubiquitous <Enter> on a computer keyboard to launch the appropriate software algorithm and have it pass through the microcosm data.

By construction, all such summary indicators would be coherent because they would be derived from the identical underlying micro data base. The summary indicators would not obscure the population's variety and heterogeneity, because the underlying micro data base would always be open (at the click of a mouse button, say, in terms of contemporary informatics functionality) for detailed inspection.

The main question is from where would this microcosm come? For the very practical reasons of cost, respondent burden, and concerns for individuals' privacy, it could not come from an omnibus longitudinal household survey. Moreover, there is not time to wait half a century or more for such a longitudinal survey to be substantially completed, by which time many things will have likely changed dramatically. The unavoidable conclusion is that the microcosm will have to be synthesized.

Such synthesis would be an extension of the synthesis of a population cohort already implicit in indicators such as life expectancy. It would differ methodologically, because the semi-aggregate or cell-based approach inherent in the underlying life table is incompatible with explicit micro data foundations. Instead, microsimulation is required.

In effect, what is being proposed is a weaving together of the ideas in Stone's SSDS (UN, 1975), with the idea of explicit integrated micro data bases proposed by a subsequent international expert group (UN, 1979). The first step is the recognition that the SSDS implicitly rests on longitudinal micro data. Indeed, Stone (1973) notes that,

> "Of course, if statistics are collected by means of a linked system of compatible records or, better still, by a continuously updated, comprehensive system of individual data *(i.e. longitudinal micro data)*, a discussion of sequence *(i.e. representations in terms of discrete time, first order Markov chains)* becomes largely irrelevant since the information in a vast, computerized data bank can be combined in any desired manner. But while these may be the methods of statistical collection in the future, they are not, with very limited exceptions, in operation at present, and so it makes sense to discuss the systematization of social statistics in terms of more familiar methods of collection." (p152, italics added)

In this sense, the future has arrived, so Stone's matrix algebra, restrictive first order Markov assumptions, and "familiar methods of data collection" need no longer be constraining.

The second step is extension of the ideas of creating integrated data bases (IDBs) synthetically, by means of statistical matching methods, so clearly articulated almost two decades ago by the UN's IDB working group (UN, 1979). They recognized the great utility of more highly multivariate micro data, as well as the practical limitations of collecting such data directly. As a result, they recommended that the desired micro data be synthesized, even if it meant that the underlying micro data records were artificial. The IDBs in that earlier work referred generally to *cross-sectional* micro data.

The bridge between these two broad ideas -- a Stone/SSDS-style framework based on longitudinal dynamics, and synthetic statistical matching of micro data -- is synthetic longitudinal micro data. The difference is that creation of synthetic longitudinal micro data requires more than techniques of statistical matching, since this matching idea does not carry over well to combining disjoint longitudinal micro data sets. The synthetic longitudinal micro data must instead be generated by dynamic microsimulation modeling (again not a new idea; see Ruggles, 1981). In essence, what is being "matched" across longitudinal micro data sets by microsimulation is not the character of individual observations, but rather observed patterns of dynamic behaviour for groups of observations in each micro data set (e.g. as sketched in a later section).

Furthermore, the synthesis of the microcosm using microsimulation means that the marginal cost of developing a "what if" capacity is negligible. For example, once a life table has been constructed, relatively little extra work is required to compute cause-deleted life expectancy. The analogous situation applies to a microsimulation basis for constructing the population microcosm. Once the investment has been made in the capacity to synthesize a "baseline" microcosm, synthesis of "variant" microcosms is relatively straightforward.

Finally, as will become evident in the description to follow, this lifecycle microanalytic approach means that one need no longer be faced with a choice between time-based and demographic styles of social accounting as discussed in Juster and Land (1981). The approach being developed here encompasses both.

3. IMPLICATIONS FOR DATA COLLECTION SYSTEMS

Consideration of a socio-economic statistical framework along the lines sketched above has major implications for both conceptual and operational aspects of data collection systems in a national statistical agency. These implications may not be that costly (relative to primary data collection costs), and most are relatively straightforward:

- data collection processes cannot exist as "stovepipe" systems, in isolation from one another;
- one kind of coordination across data collection processes is use of common concepts and definitions (e.g. identical definitions and methods for eliciting educational attainment);

- the other kind of coordination is assuring appropriate overlap in content -- basically to anticipate the need for synthetic statistical matching (or equivalent methodologies); and
- microanalytic uses of raw data are far more demanding of data quality than aggregative uses.

In effect, this means that data collection systems must be jointly planned, and that micro level data quality standards must be more stringent.

The joint planning requirement is not new. Construction of the SNA also requires some coordination of data feeder systems, not least to assure that there is some method to cover all sectors of the economy. However, this coordination is much less onerous than that entailed by microsimulation. The reason is that inconsistencies across data collection systems uncovered by the SNA can be resolved at the high level of "macro editing". Adjustments are made to broad aggregates, notwithstanding the fact that this introduces inconsistencies between various SNA aggregates and their source micro data. However, for the microsimulation purposes here, a key objective is internal consistency across source data sets at the micro level.

The micro level data quality requirement is also not new. It has been faced most acutely whenever a public use micro data set is produced. Knowing that users will subject the data to intensive inspection and analysis, for example as part of assessing regression "outliers", extensive editing and imputation is applied to these data. Similar but weaker micro level data quality concerns are faced with population census micro data files that, while not publicly available, are open to generalised ad hoc cross-tabulation requests.

Still, micro level data quality concerns will be much more acute in the context of an integrative microanalytic framework such as that about to be described. It is one thing for a process of edit and imputation to assure that each record in a given micro data set is plausible and internally consistent. It is quite another to assure that multiple micro data sets are mutually consistent -- for example that a health survey and a disability survey yield the same age- and sex-specific distributions of disability by severity, or that a longitudinal survey on labour dynamics produces cross-sectional estimates of labour force participation that agree with those generated by the mainline labour force survey, or that a time series of administrative data on school enrollments is consistent with census data on educational attainment by age and sex.

This requirement for mutual consistency highlights a concern raised by Wilk (1987), namely the relative weakness of statistical methods for addressing non-sampling error. For example, item non-response or bias in household surveys typically causes serious under-reporting of selected income sources. However, conventional edit and imputation processes usually address this in only a limited manner (e.g. income components falsely reported as zero are not changed). It is only the community of microsimulation modelers for tax/transfer policy who have, of necessity, had to grapple with this problem (Citro and Hanushek, 1991; Bordt et. al., 1990; Wolfson et al., 1989). Moreover, household survey editing does virtually nothing about response rounding (e.g. giving income to the nearest $100 or $1000) -- even though there is evidence that such respondent behaviour can cause errors in some statistics (e.g. quantiles) of the same magnitude as conventional sampling error (Rowe and Gribble, 1994). Finally, there is a growing recognition of the importance of longitudinal surveys, which are clearly fundamental to developing descriptions of dynamics, and to disentangling causal pathways. Using longitudinal micro data for these purposes will entail the use of more sophisticated inferential methods than statistical agencies typically encounter, for example hazard regression as compared to cross tabulation. This in turn should expose the data to far more critical scrutiny.

4. THE LIFEPATHS PROJECT

We turn now to an illustration of these general points. The LifePaths project is an effort to construct a prototype socio-economic statistical framework. The project is being undertaken by Statistics Canada on behalf of the Canadian Ministry of Human Resources Development, the recently created "super-ministry" responsible for welfare, pensions, unemployment insurance, and labour market policies, among others.

The basic objective of the LifePaths statistical framework is to provide a coherent and multi-faceted series of "views" of the socio-economic status of the Canadian population. This framework is designed to have the general characteristics indicated earlier, namely a capacity to indicate overall outcomes and variety, and to provide answers to "what if" questions. The substantive domain includes how Canadians are spending their time in various activities such as working, learning, family roles, participating in government programs, and leisure.

Generalizations of working life tables form one of the central views or facets to be provided. Table 1, for example, shows for Canadian male birth cohorts, not only conventional life expectancies, but also the average ages at which men could expect their first entry and last exit from the paid labour force. By examining a series of these (period) birth cohorts, each representing a successive decade, the analysis vividly displays the long run trends of more time spent in schooling, ever earlier ages of retirement, and a general reduction in working years for men. The final column also gives a clear indication of the impacts of these trends on public pension costs. (Note that while old, these are apparently the most recent working life table estimates available.)

The LifePaths framework extends these basic working life table results in several directions. Annual work patterns are considered in more detail, going beyond a two-way breakdown between working and non-working years. For example, part-time work, increased duration of paid holidays and vacations, changes in typical hours worked per week, sub-annual spells of unemployment or withdrawal from the labour force, periods where work and schooling are simultaneously pursued, and more participation in self-employment are all taken into account. In addition, the time aspects of work are combined with the economic aspects, particularly income.

Other major forms of activity are also included. One is formal schooling; another is familial context (e.g. living alone or with other family members). Thus, involvements with the major institutions of society -- work, school, and family -- are covered. The LifePaths framework therefore combines both the "active" (learning and earning) and "passive" sequences ("succession of family groupings to which individuals are attached in the course of their life", p145) in Stone's (1973) demographic accounting SSDS proposal. This is a capacity that in practice becomes combinatorially intractable with the matrix methods he used.

Additionally, participation in major social programs is planned -- for example, Social Assistance (SA), Unemployment Insurance (UI), and Workers Compensation (WC) disability pensions. Generally, a more fine-grained account of time use is included, based on data from time use surveys -- the time-based accounting proposed, for example, by Juster et. al. (1981). Thus, major categories of activity include not only work and school, but also unpaid housework, personal care, care for others, sleep, commuting, TV, other passive leisure, active leisure, interaction with family members, and other social interactions.

The LifePaths framework encompasses all these human activities, from a complete life cycle perspective, in a coherent and integrated manner -- thereby combining and nesting both time-based and demographic accounting approaches as debated in Juster and Land (1981). Constructing the LifePaths framework is challenging, and the results to be presented here are a prototype.

Methodologically, the LifePaths framework is premised on several major statistical innovations.

First, no single data set contains all the required information, for example detailed data on human activities from both economic and social perspectives.

Table 1 -- Historical Stationary Male Life and Working Life Expectancies at Age 15

	average age at			number of		
Year	entry to labour force	retirement	dealth	working years	retirement years	working years per year of retirement
1921	16.5	63.7	67.6	47.2	3.9	12.1
1931	17.0	64.0	68.4	47.0	4.4	10.7
1941	17.2	64.1	69.1	46.9	5.0	9.4
1951	17.5	63.9	70.4	46.4	6.5	7.1
1961	18.2	64.0	71.2	45.8	7.2	6.4
1971	19.8	63.3	71.3	43.5	8.0	5.4

Source: Gnanasekaran and Montigny (1975) and Wolfson (1979)

Current and planned data sets in this domain are partial and fragmentary. Moreover, as already noted, considerations of cost, respondent burden, and privacy suggest that fully integrated household survey data will never be practical. Thus, processes of synthetic integration, utilizing multiple data sets, are inevitably required.

Second, the framework is intended to cover individuals' full life cycle histories. Doing so with actual longitudinal data would require decades of survey follow-up, by which time many things will have changed. Thus, the basic idea is to build on and generalize the concept of period life expectancy and its underlying life table. In turn, this means that the analysis will focus on realistic, but hypothetical, population cohorts.

A third part of the basic objective is the detailed reflection of individual variety or heterogeneity, and in turn, a capacity to view distributional phenomena such as income inequality. This capacity requires explicit micro data foundations. Since data on the actual life paths of a representative sample of individuals is infeasible, the underlying micro data must be synthetic. Yet at the same time, these data must be sufficiently realistic to be essentially indistinguishable from the partial sets of characteristics observed in real data from actual population samples, including longitudinal surveys.

These requirements imply that the heart of the LifePaths statistical framework must be a microsimulation model. In other words, the core of the statistical framework is a sample of realistic, but synthetic, individual life paths.

5. ON SYNTHETIC DATA

Before presenting initial results, it is important to explain the sense in which the LifePaths framework is based on synthetic data, and the extent to which these synthetic results are a reasonable reflection of current realities.

Human lifetimes typically span about three-quarters of a century. However, given the relatively rapid pace of change over a wide range of human activities, it is almost impossible to have consistent and stable socio-economic observations for this length of time. Statistics that have been well accepted for decades simply did not exist 75 years ago -- for example the unemployment rate, GDP per capita, and measures of leisure time. Correspondingly, it is quite possible that 75 years from now, in 2070, these basic statistics, whose importance is taken for granted today, will have been superseded by new kinds of statistics we can barely imagine.

Yet there is a very broad interest in statistical indicators that do reflect processes spanning a human lifetime. The most obvious is life expectancy. Other such statistics are the proportions of marriages that can be expected to end in divorce, the number of different jobs an individual can expect to have over his or her working career, the expected adequacy of public pensions relative to pre-retirement earnings, and the portion of life expectancy that will be spent in good or ill health. Clearly, such lifetime statistical indicators exist and are more or less widely accepted. The LifePaths framework generalizes such indicators.

While it may not be widely appreciated, life expectancy is a "made up" statistic. It is analogous to a statement about where a car is heading based on its position and velocity while ignoring any acceleration. Life expectancy is based on current age- (and sex-) specific death rates so, like vehicle speed, it is based on real data. But (period) life expectancy applies to a hypothetical individual who has been taken out of calendar time, and spends his or her entire lifetime exposed to the mortality rates of the early 1990s. In essence, any acceleration or deceleration of mortality rates is ignored.

It is, of course, well known that mortality rates have generally fallen over past decades, and it is widely anticipated that these declines will continue. Thus, while life expectancy itself ignores these trends in mortality rates, trends in life expectancy provide very convenient summary indicators of these changes in underlying mortality rates, since they track a form of weighted average of age- (and sex-) specific mortality rates, all of which are changing over time. The underlying age-specific mortality rates are always available for inspection, but trying to make sense of the evolution of even one hundred numbers is a complex task. (More numbers are involved if mortality is broken down by sex and marital status as well as by single year of age.) Life expectancy is a helpful indicator precisely because it collapses these hundred numbers into a single intuitively accessible indicator -- one whose changes over time correspond reasonably to the changes over time in the underlying age-specific mortality rates.

The LifePaths framework is designed to be completely analogous. However, it builds on a much richer variety of processes and statistical descriptions of individuals' transitions among various life states. For example, in addition to mortality, explicit account is taken of demographic states such as marital status, and the associated transitions of entering a common law or legal union, and leaving a union to a separated or divorced state. Similarly, other socio-economic status classifications like working or engaging in learning have been included, based on real data on recent distributions and transition rates amongst these states.

In order to achieve this kind of generalization of life expectancy, the underlying concept of a life table has had to be generalized as well. In a life table, the finest level of detail is a group of individuals -- for example as defined by sex and single year of age. Within such a group, all individuals are assumed to be homogeneous. In LifePaths, this level of detail is insufficient. Explicit consideration of heterogeneous individuals characterized by multiple attributes is essential in order to make the best use of, and to reflect most accurately, results emerging from analyses of dynamic behaviour patterns in rich longitudinal micro data sets.

In an important sense, this implies that LifePaths produces results that are much more realistic than life expectancy produced from a conventional life table. For example, in LifePaths, mortality rates are broken down by marital status as well as age and sex; and in turn marital status depends in a complex way on factors like educational attainment, fertility history, and labour force activity durations.

On the other hand, the synthetic character of the "data" underlying LifePath results is inevitably more explicit than in the case of a life table. While a population of individuals underlies any conventional life table, the individuals themselves are only implicit -- all that is calculated is the numbers of individuals in each cell or category, e.g. by age and sex. In contrast, in the LifePaths framework, all individual life paths must be explicit.

So what meaning should be attached to a LifePath result such as a breakdown of life expectancy into the number of years an individual can expect to spend in the paid labour force and in learning? The interpretation should be analogous to conventional life expectancy -- a kind of summary of recent population flow rates. LifePaths results show how things would be if recent rates of transitions among socio-economic states (conditional on attributes for heterogeneous individuals) were constant.

6. INITIAL RESULTS

The LifePaths statistical framework consists, fundamentally, of a sample of complete (synthetic) individual life cycle histories. However, this longitudinal micro data base of sampled life histories is far too complex to be examined directly, so we offer here only a few summary "views" of the underlying microcosm. The specific views start from conventional demographic analysis.

Note that these "views" stop short of summary scalar indicators like GDP; they show simultaneously a number of basic population attributes. This need not be seen as a weakness, as the inability to make the last step to a single overall measure as in the SNA. Rather, a given "view" can be seen as a demonstration of the power of contemporary computer graphics to facilitate more textured appreciations of social economies than is possible with a single index.

To start, one of the most basic demographic images is the population pyramid. Figure 1 shows such a pyramid for the base case life table population, with counts of females along the horizontal axis to the right, male counts to the left, and age to 100 along the common vertical axis. It is based on period (late 1980s and early 1990s) transition probability functions, which are sketched in a later section. As expected, at higher ages, the survival curve for females falls more slowly than that for males, a counterpart to (or more accurately the underlying reason for) females' higher life expectancy. (The blip in the age 99 interval reflects the fact that this is actually the age ≥99 interval.)

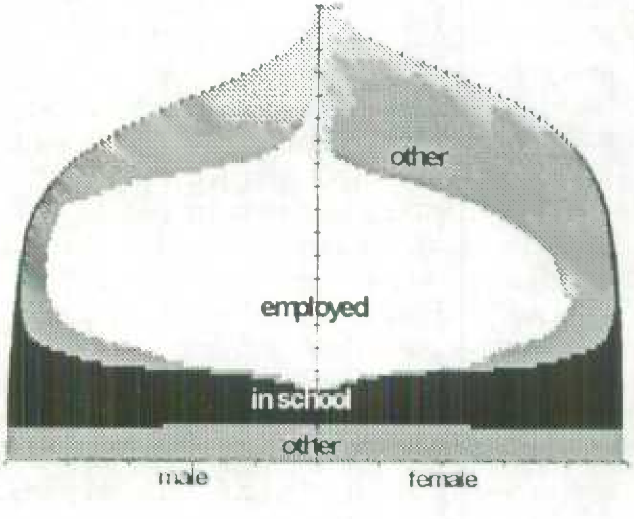


Figure 1 -- LifePaths Population (person-years) by Major Activity, Age and Sex

Figure 1 also shows the population broken down into three socio-economic categories -- "employed", in "school", and "other". "School" starts at grade 1, so daycare and kindergarten are part of "other". Since the LifePaths framework tracks individuals through time continuously, some arbitrary decisions have been applied in years where individuals engage in more than one activity. Specifically, to be considered "employed" in this diagram, the individual had to be working at least 15 hours per week, and the plurality of time during the year had to be spent working at this rate. Thus, someone who spent 5 months as a student, 4 months working at least 15 hours per week, and the remaining 3 months of the year working less than 15 hours per week (including not working at all) would be considered in "school" that year; while if the 5 and 4 were reversed, they would be considered "employed". (Definitions such as these are under the control of the LifePaths user.) The diagram shows that virtually everyone is in school by age 8, a few start leaving at age 16, most have left by age 20, but there is a tail of both males and females who are in school through their twenties.

No one appears to make a transition directly from school to employment, though we return to this point in a later figure. Instead, perhaps a surprising proportion of individuals are in the "other" category, which includes the unemployed as well as those not in the labour force (e.g. homemakers, the retired). As expected, males are more likely to be employed at various ages than are females. The employed portion of the population shows a dip in the age-related trend to higher participation for women in the prime child-bearing years 20-25, and then something of an acceleration in the 25-35 age range. Men show a relatively sharp decline in participation in the age 60-65 age range.

Figure1 corresponds to Stone's "active sequence" (i.e. transitions among working and learning states), while Figure 2 gives an overview of his "passive sequence". It uses the same population pyramid graphic form, and refers to exactly the same underlying LifePaths synthetic population, but classifies individuals along a different dimension, family status. By definition, all individuals under age 18 are classified as "growing up" unless they are married or have a child. Also, whenever a marriage breaks down, any children are assumed to remain with the mother. This assumption explains why there are female, but no male lone parents. (Future versions will incorporate more realistic data on custody arrangements.)

Comparing the male and female curves for the married states (couples with and without children) shows the male curves displaced a few years toward higher ages. This is a reflection of the general pattern where husbands tend to be a few years older than their wives. The diagram also shows there are many more widows than widowers. This is a consequence of both the positive average age difference between husbands and wives, and the greater life expectancy of women. Finally, the diagram indicates the much higher rates of institutionalization of women (principally in nursing or chronic care facilities), due in turn to their greater longevity and higher prevalence of health problems at older ages, and the fact that similarly incapacitated males more often have a wife who can care for them at home.

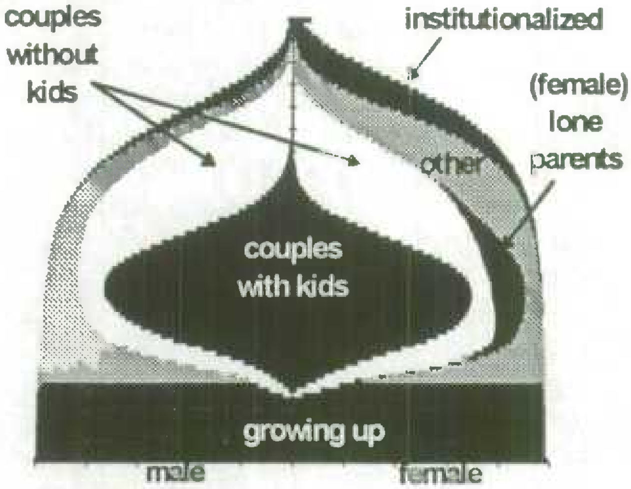


Figure 2 -- LifePaths Population (person-years) by Family Status, Age, and Sex

Figures 1 and 2 show only the beginnings of the LifePaths framework; they are simply two "views" (in this case cross-tabulations) of the full underlying microcosm -- a longitudinal micro data set for a synthetic "early 1990s" period birth cohort. Exactly this same underlying longitudinal micro data set can be tabulated to generate the view in Figure 3, which shows flows between states rather than stocks of individuals within each state. In this case, Figure 3 graphs the flows corresponding to the stocks in Figure 1. The horizontal axis shows the number of individuals making each kind of transition each year, again in population pyramid style with age along the common vertical axis, females on the right horizontal axis, and males on the left. (The extremes of the horizontal axis span 18% of the population, so that for a cohort of 100,000 the maximal male and female flows shown are each 9,000 per year.)

The first transition is from "other" (early childhood or pre-school) to "school". Figure 1 indicates that all male and female children make this transition by ages 6 and 7. The next major transition is at the end of "school", where the peak flow rate to "employed" occurs around age 20 for both males and females. A smaller number, also peaking at about age 20, move from school to "other" activity. Recall that the "other" category is any person-year where the plurality of the year (i.e. at least a tiny bit more than one-third) was spent neither as a student nor working more than 15 hours per week.

From early adult ages to the 60s, the main flows are between the "employed" and "other" categories. Note that all these flows are gross rather than net. It is notable that the net flow between employed and other (based on comparing the gross flows) shifts direction toward "other" in the 40-45 age range for females, but remains quite small for males through age 50. This is followed by retirement peaks in the 55- 65 age range, the one for males being more pronounced.

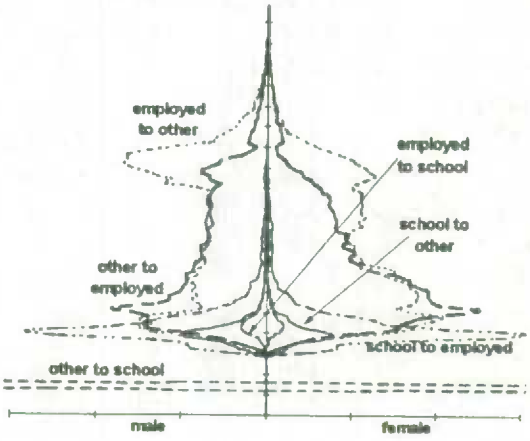


Figure 3 -- LifePaths Gross Flows Between Major Activities (persons per year) by Age and Sex

In addition to stocks and flows of individuals in various categories of activity, the LifePaths framework also supports data views showing sojourn times -- lengths of time individuals spend in various states. Such sojourn times have already been illustrated in Table 1 above, giving earlier estimates of working life expectancy. A major additional capability in LifePaths, given its explicit micro data foundations, is views of uni- or bivariate distributions of durations or sojourn times across the population -- for example the joint distribution of years of school and employment for males and females. (Space limits preclude showing any of these graphs.)

Figure 4 gives one more image from the basic LifePaths simulation -- but this time showing another classification of activities, and using a different horizontal axis. Instead of person-years from a period life table birth cohort as in Figures 1 and 2, in Figure 4 the horizontal axis shows major activities in terms of the number of hours spent in each activity during an average week (i.e. 168 hours) -- for each sex and single year of age. Thus, for example, market work here is shown averaging about 40 hours per week for males age 30 to 55.

Superficially, Figure 4 looks exactly like data that could be produced directly from a time use survey, and as a matter of validation, it should be very close. However, it was generated by the LifePaths simulation, and differs somewhat from the underlying time use survey data principally because the data have been made coherent. For example, the underlying annual labour force participation rates by age and sex in Figure 4 are consistent with those underlying Figure 1, and the demographic patterns with Figure 2 -- by construction.

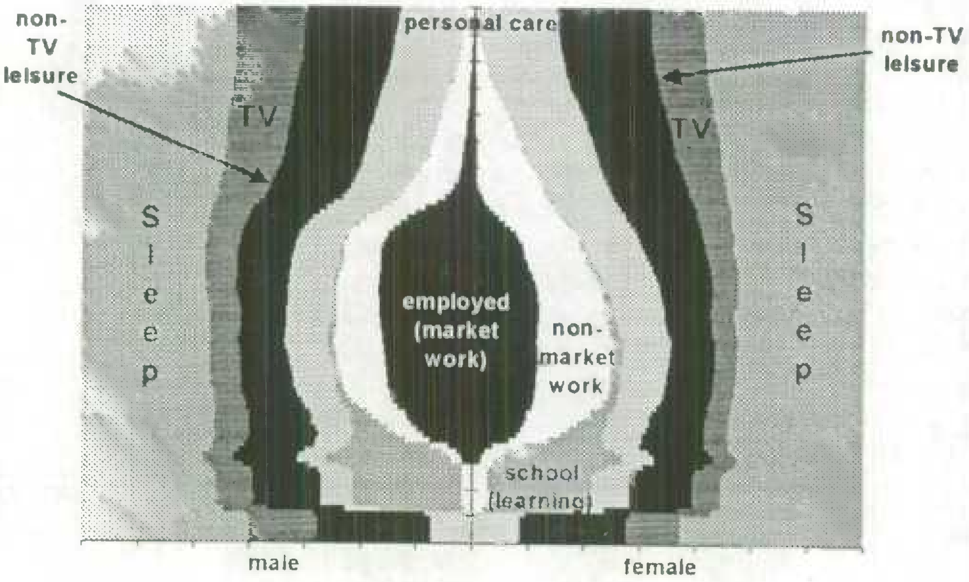


Figure 4 -- LifePaths Time Use (average hours per week) by Major Activity, Age and Sex

One impression left by the diagram is the relatively small proportion of average male and female lifetimes spent in "market work" -- the ostensible domain of the SNA. When viewed from the perspective of average weekly hours (rather than whether more than a third of the year was spent working more than 15 hours per week, as in Figure 1), market work is a very small portion of a total (or even waking) lifetime. Of course, non-market work and the consumption aspects of personal care and use of leisure time also have important economic aspects, but they are not captured in the SNA beyond aggregate dollar measures of personal consumption by commodity.

This figure also indicates the limitations of conventional demographic dependency ratios -- which use raw counts of individuals of working age (e.g. age 20 to 64) as the denominator. In the context of Figure 4, such ratios clearly understate the degree of economic dependence of many individuals in society. The diagram also suggests the need to represent more explicitly the mechanisms by which purchasing power, generated principally by time spent in "market work", is made available to the rest of the population. These mechanisms include intra-family transfers, and government tax/transfer programs. More generally, this combination of a time use with the more conventional demographic framework in LifePaths offers the opportunity to construct a coherent series of statistical views that provide a much more comprehensive accounting of social and economic activity.

LifePaths images such as Figure 4 clearly show there is much more to life than is captured in the market economy focus of the SNA. It follows that regular publication of this kind of statistical framework could have an important effect on public policy discussion. It would place economic factors in a broader context, and draw attention to a much wider range of impacts of policies directed toward unemployment, retirement, income redistribution, education, childcare, de-institutionalization, and the work week -- to name a few.

It should be emphasized again that these results from the LifePaths statistical framework are still substantially illustrative. The underlying synthetic longitudinal micro data set is still under development. As will be described in the next section, these underlying data are based on a range of recent surveys and analyses -- i.e. real data. But the underlying analyses in part still involve preliminary results.

7. UNDERLYING METHODS

The LifePaths framework just illustrated draws particularly on two recent data sets, and almost a decade of development of related microsimulation models. The recent data sets are the 1992 General Social Survey (GSS), which includes detailed questions on time use based on 24 hour recall, and the Labour Market Activities Survey (LMAS), which provides detailed longitudinal data on labour market dynamics over the 1988 to 1990 period. The LifePaths microsimulation model in turn is a combination of the results of the GSS and LMAS analyses, the DEMOGEN microsimulation model (Wolfson, 1989) as it has recently been re-implemented in the newly created ModGen C++ microsimulation software environment, and the new post-secondary education Income Contingent Repayment Loan (ICL) model being developed for the Human Resources Development Ministry of the Government of Canada.

This section gives a very brief overview of the processes involved in synthesizing a LifePaths birth cohort, the core of the LifePaths statistical framework. Generally, the synthesis process involves an overall architecture connecting a series of economic and socio-demographic processes, and detailed data analysis to develop empirically based statistical descriptions of each process (i.e. behaviour dynamics).

As in a conventional life table, LifePaths starts with a specified population of individuals, say 100,000 births. Unlike a life table, however, each individual is followed over time until his or her death. (A life table, in contrast, follows groups of individuals, all of whom are considered homogeneous.) At any moment in time, an individual faces a chance of making a transition. Depending on his or her current state or set of attributes, this could be a transition into the labour force, or into a marital union. Which transitions are possible depends on the range of states that are explicitly considered. In the current version of LifePaths, individuals are jointly characterized by the following basic attributes at each point in their lives:

- age -- as a continuous variable
- fertility -- ages at the birth of children, presence of children in the familial home
- nuptiality -- unattached, in a common-law or marital union, separated, or divorced
- work status -- including labour force participation and employment status (hours per week, weeks in the year)
- school status -- grade and type of institution if attending, educational attainment

- work income -- hourly rate, weekly and annual earnings
- time use -- categories shown in Figure 4 plus finer disaggregations
- program participation -- including welfare, unemployment insurance, public pensions
- spouse attributes -- including age, educational attainment, labour market experience

In addition, a wide range of derived attributes can be constructed from these basic attributes such as the variables shown in Figures 1 to 4.

Given this listing of attributes, the next step in describing LifePaths is the processes by which the trajectories for each attribute is generated. A brief sketch is given in the following paragraphs.

Demography -- Fertility is modeled as the sequel to conception, which in turn is modeled as a series of piecewise constant hazard rates, conditional on age, marital status, and number of previous live births. The main data source is birth registrations, supplemented by data from the 1983 Family History Survey to account for biases arising from conception while single or in a common law union, followed by marriage before the birth of the child. Mortality rates are conditional on age, sex and marital status, and are based on death registrations. In both cases, the population census provides the denominators.

Union formation and dissolution are represented by a series of hazard functions. From the single state, there are competing risks of entering a common-law union or a legal marriage. Marriage breakdown involves risks of separation and subsequent divorce. These hazards have been separately estimated for men and women, and depend in a complex way on previous history. For example, females' "risk" of entry to a union is positively related to being pregnant, and is highest shortly following labour force entry. Risks of separation for females are higher if there are no young children at home, if the woman was a teenage bride, and if the woman has recent work experience.

Educational Progression -- Transition rates for progression through elementary and secondary school were constructed to be as close to jointly consistent as possible with the 1986 and 1991 population census data on the school attendance rates of children of the relevant ages. Progression through post-secondary institutions (colleges, trade schools, universities) is based on hazard rates jointly estimated from the National Graduates Survey (NGS), administrative data on school enrollments, and the Labour Market Activities Survey (LMAS) for cases where young people quit work to return to and continue their studies.

Labour Market -- Labour market experience is simulated in two main parts -- whether or not employed, and earnings from employment. The first of these, transitions into and out of employment, is estimated from the LMAS separately for males and females, and also separately for first entry, second and subsequent entry, and exit from employment. First entry is represented by waiting time distributions, while the other transitions are represented by multivariate hazard functions. Sex and educational attainment are important determinants of the waiting time to first employment. Re-entry hazards depend on sex, educational attainment, and duration of the current spell of non-employment, and for women the presence of infant children has an additional depressing effect.

Earnings are in turn based on employment status as just described, and separate models for weekly hours of work, and hourly wages. Upon first entry to employment, a weekly hours value is randomly assigned drawn from an age-, sex- and educational attainment-specific distribution, in turn based on data from a combination of the NGS, LMAS, and the Survey of Consumer Finances (SCF -- the annual household income distribution survey). Subsequently, the weekly hours variable is updated as a function of age, sex, last year's weekly hours, and educational attainment. At the same time that weekly hours is assigned, each individual is assigned a percentile rank for hourly earnings. The hourly earnings rate is then "looked up" from age-, sex- and educational attainment-specific distributions. Percentile ranks are adjusted from year to year based on estimates of rank order "churning" from the LMAS.

Daily Time Use -- The 1992 General Social Survey (GSS) collected 24 hour time use diary data for about 9,000 individuals, evenly distributed by age, sex, day of the week, and month of the year. The GSS also collected basic data on educational attainment, employment status, and family status. After extensive analysis of these data, a LifePaths module was created which imputes to every simulated person-day one vector of time spent over a 24 hour period in each of a series of activities, including at the highest level of aggregation the categories shown in Figure 4. (Special assumptions have been made for children under age 15 and those elderly living in institutions, since they were not covered by the GSS.)

The statistical analysis indicated that age, sex, day of the week, marital status, presence of young children, educational attainment, and main activity (i.e. student, employed or self-employed, other) were all significantly associated with these vector patterns. Thus, all of these attributes, as generated by other LifePaths processes, were used in the imputation. The imputation process the observed variability in time use patterns amongst individuals with the same attributes, essentially by using the distribution of vector residuals from a multivariate regression analysis.

8. VALIDATION AND DATA QUALITY CONCERNS

Validating the LifePaths model is fundamentally impossible. The reason, simply, is that its intent is to create an instance of a sample from a hypothetical birth cohort. Thus, no comparison with "reality" is ever possible. However, the synthetic microcosm of individual life paths should, by construction, reproduce the major marginal joint distributions from which it was built -- for example, labour force participation rates , fertility rates, mortality rates, union formation and dissolution rates, educational enrollment rates, and age/sex-specific distributions of labour market earnings.

During the course of constructing the LifePaths prototype described in this paper, all these comparisons have been continually checked. By and large, agreement is good. The main instances of disagreement arise when the underlying data sources are not themselves consistent with each other. If anything, this is a signal of error in the source data. In effect, LifePaths has provided a framework for socio-economic micro data, in part analogous to the SNA framework, wherein data from diverse sources are rendered coherent, and inconsistencies thereby highlighted.

9. CONCLUDING COMMENTS

This paper started with users' needs for more comprehensive and coherent socio-economic statistical information, and offered a diagnosis of the failures of earlier international efforts to address these needs. A new approach is suggested, based on much more extensive use of a range of multivariate micro data sets, and microsimulation methods. Initial features of the approach -- particularly comprehensiveness and coherence -- have been illustrated with preliminary results from the LifePaths model under development at Statistics Canada.

Space has not permitted other features to be graphically illustrated, such as the explicit micro data foundations and hence the capacity to display variety. Further work is required to illustrate other key features such as summary indicators (e.g. lifetime income distributions), and "what if" simulations. Still, the results presented constitute a substantial "proof by construction" of the practical and technical feasibility of the approach.

At the same time, the approach highlights gaps and weaknesses in existing socio-economic statistical data, particularly from a microanalytic perspective. The LifePaths approach would place much stronger demands on the coherence and quality of underlying socio-economic surveys and data collection systems. Given a measure of acceptance of the benefits for socio-economic statistical reporting of something like the LifePaths approach, it can provide the basis for strategic planning in national statistical agencies.

REFERENCES

Bordt, M., Cameron, G., Gribble, S., Murphy, B., Rowe, G., and Wolfson, M. (1990). The social policy simulation database and model: An integrated tool for tax/transfer policy analysis, *Canadian Tax Journal*, 38:48-65.

Citro, C.F., and Hanushek, E.A. (1991). *Improving Information for Social Policy Decisions, The Uses of Microsimulation Modeling,* National Academy Press, Washington, D.C.

Easton, G.S., and McCulloch, R.E. (1990). A multivariate generalization of quantile-quantile plots", *Journal of the American Statistical Association*, June, Vol. 88, No. 410, Theory and Methods, 376-386.

Garonna, P. (1994). Statistics facing the concerns of a changing society, *Statistical Journal of the United Nations ECE*, Vol 11, No. 2, 147-156.

Gnanasekaran, K.S., and Montigny G. (1975). *Working life tables for males in Canada and Provinces, 1971*, Statistics Canada Catalogue 71-524E Occassional, Ottawa.

Juster, F.T., and Land, K.C. (1981). Social accounting systems: An overview, in F.T.Juster and K.C.Land (Eds), *Social Accounting Systems -- Essays in the State of the Art*, Academic Press, New York.

Juster, F.T., Courant, P.N., and Dow, G.K. (1981). The theory and measurement of Well-Being: A suggested framework for accounting and analysis, in F.T.Juster and K.C.Land (Eds), *Social Accounting Systems -- Essays in the State of the Art,* Academic Press, New York.

Mathers, C., and Robine, J-M (1993). Health expectancy indicators: A review of the work of REVES to date, in J-M Robine, C.D.Mathers, M.B.Bone, I.Romieu (Eds), *Calculation of Health Expectancies: Harmonization, Consensus Achieved and Future Perspectives*, INSERM / John Libby Eurotext Ltd., Vol. 226.

Moser, Sir C. (1973). Social Indicators -- Systems, Methods and Problems, *Review of Income and Wealth*, Series 19, No.2, June, 133-141.

OECD (1976). *Measuring Social Well-Being*, Paris.

OECD (1977). Basic Disaggregations of Main Social Indicators, D.F.Johnston, *Special Studies No. 4, The OECD Social Indicator Development Programme*, Paris.

OECD (1982). *The OECD List of Social Indicators*, Paris.

Pommier, P. (1981). Social Expenditure: Socialization Expenditure? The French Experience with Satellite Accounts, *Review of Income and Wealth*, December.

Pyatt (1990). Accounting for Time Use, *Review of Income and Wealth*, Series 36, No. 1, March, 33-52.

Rowe, G., and Gribble, S. (1994). Income statistics from survey data: Effects of respondent rounding, forthcoming in Proceedings of the American Statistical Association, Section on Government Statistics.

Ruggles, N., and Ruggles, R. (1973). A proposal for a system of economic and social accounts, in M. Moss (ed.), *The Measurement of Economic and Social Performance*, National Bureau of Economic Research, New York.

Ruggles, R. (1981). The Conceptual and Empirical Strengths and Limitations of Demographic and Time-Based Accounts, in F.T.Juster and K.C.Land (Eds), *Social Accounting Systems -- Essays in the State of the Art,* Academic Press, New York.

Stone, R. (1973). A system of social matrices, *Review of Income and Wealth*, Series 19, No.2, June, 143-166.

United Nations (1975). *Towards a System of Social and Demographic Statistics* (SSDS), Studies in Methods, Series F, No. 18, ST/ESA/STAT/SER F/18, New York.

United Nations (1979). *The Development of Integrated Data Bases for Social, Economic, and Demographic Statistics* (IDBs), Studies in Methods, Series F, No. 27, ST/ESA/STAT/SER F/27, New York.

Vanoli, A. (1994). Extension of National Accounts: opportunities provided by the implementation of the 1993 SNA, *Statistical Journal of the United Nations ECE,* Vol 11, No. 3, 183-191.

Wilk, M.B. (1987). The Concept of Error in Statistical and Scientific Work, paper presented to the U.S. Bureau of the Census Third Annual Research Conference, Baltimore.

Wolfson, M.C. (1979). Saving for Retirement: How Much is Required, Volume II, Appendix 18 in *The Retirement Income System in Canada: Problems and Alternative Policies for Reform,* Task Force on Retirement Income Policy, Ministry of Finance, Ottawa.

Wolfson, M.C. (1989). Divorce, Homemaker Pensions, and Lifecycle Analysis, *Population Research and Policy Review*, 8: 25-54.

Wolfson, M.C., Gribble, S., Bordt, M., Murphy, B., and Rowe, G. (1989). The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration, *Survey of Current Business*, 69, 36-40.

Wolfson, M.C. (1994). Implications of Evolutionary Economics for Measurement in the SNA, Towards a System of Social and Economic Statistics, paper prepared for the Twenty-Third General Conference of the International Association for Research in Income and Wealth St. Andrews, New Brunswick, August, 21-27, 1994, mimeo, Statistics Canada, Ottawa.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

DEVELOPMENT, USE AND MODIFICATION OF HEALTH RISK APPRAISAL FUNCTIONS: THE FRAMINGHAM STUDY

R.B. D'Agostino[1]

ABSTRACT

Health Risk Appraisal functions are mathematical functions or models which relate risk factor variables to the probability of developing an event such as coronary heart disease. The Framingham Study has been a leader in developing functions for cardiovascular diseases. This article reviews the history of the development and some of the uses of these functions. It also presents some recent modifications to accommodate specific mathematical and practical concerns.

KEY WORDS: Prediction models; Risk factors; Risk profiles; Epidemiological studies.

1. INTRODUCTION

The Framingham Study is an ongoing major prospective cohort epidemiological study begun in 1948 whose primary objective is to investigate and establish the relation of cardiovascular disease (CVD) to risk factors such as age, sex, blood pressure, cholesterol, cigarette smoking, hematocrit, obesity and diabetes (D'Agostino and Kannel, 1990). CVD encompasses coronary heart disease (myocardial infarction, unstable angina and stable angina), stroke, congestive heart failure, intermittent claudication, and cardiovascular and cardiac deaths. The study enrolled 5,209 subjects, 28 to 62 years of age (2336 males and 2879 females). Subjects return every two years for a physical examination and history in which the CVD risk factors are updated and information on the status and development of cardiovascular disease since the last visit are obtained. In addition, the study maintains an extensive surveillance of the subjects at all times to note and obtain information on death and the development and course of CVD.

Over the years the Framingham Study has developed **mathematical predictive models** which relate the risk factors to the probability of developing CVD events or subsets of CVD events such as coronary heart disease. The models or functions are today called **health risk appraisal functions**. In the following we present a brief history of the development and use of these functions along with some recent advances in them.

2. INITIAL YEARS (1948 - 1976)

During the first decade of the study data were accumulated and presentations in the literature focused mainly on study design and objectives (Dawber, Meadors, Moore (1951) and Dawber, Kannel, Lyell (1963)). The first paper in which predictive functions were presented appeared in the mid 1960s (Truett, Cornfield and Kannel, 1967).

Table 1 displays two of the original functions. These are Fisher's linear discriminant functions and they relate the major CVD risk factors to the development of a first coronary heart disease (CHD) event. The risk factors are: age in years (AGE), total serum cholesterol (CHOL), systolic blood pressure (SBP), metropolitan relative weight (MRW), hematocrit (HEM), cigarette smoking as a yes or no (CIG) and left ventricular hypertrophy as measured by the electrocardiogram (LVH). The variable MRW is computed by dividing a subject's actual weight to an ideal weight given by the Metropolitan Life Insurance tables of ideal weights. All the risk factors were significant at the p = 0.05 level except those identified by an (NS).

[1] Ralph B. D'Agostino, Professor of Mathematics/Statistics and Public Health, Boston University, 111 Cummington Street, Boston, MA 02215, USA

Table 1

Fisher's Linear Discriminant Functions for Relating Risk Factors to the Development of First Coronary Heart Disease (CHD) Event Within a 12 Year Period

Subjects were free of CHD at Examination 1, some of which develop CHD within 12 years of follow-up from Examination 1 (MEN: n=2187 with 258 CHDs and FEMALES: n=2669 with 129 CHDs)

	Men		Women	
	Coefficients			
Constant	-10.8986		-12.5933	
AGE	0.0708		0.0765	
CHOL	0.0105		0.0061	
SBP	0.0166		0.0221	
MRW	0.0138		0.0053	(NS)
HEM	-0.0837	(NS)	0.0355	(NS)
CIG	0.3610		0.0766	(NS)
LVH	1.0459		1.4338	

NS = Non-significant at 0.05 level. All other variables were significant at least at the 0.05 level.

These functions could be used for classification purposes as follows. Let F represent the Fisher Linear Discriminant Function defined as

$$F = A + B_1 * X_1 + B_2 * X_2 + \ldots + B_K * X_K \tag{1}$$

where $X_1, \ldots, X_K$ represent the values of the risk factors and $B_1, \ldots, B_K$ are the coefficients whose numerical values are given in Table 1. For a given individual, obtain the risks factors of Table 1 and compute the F value of equation (1). The classification rule is:

If F ≥ 0, then classify subject as CHD
If F < 0, then classify subject as non-CHD

Further, the investigators noted that the probability of developing CHD within 12 years could be estimated by exponentiating the function F of (1). In particular, this estimates the conditional probability of developing a CHD given the data consisting of the risk factors $X_1, \ldots, X_K$. In symbols this is

$$P(CHD \mid X) = [1 + \exp(-F)]^{-1} \tag{2}$$

As displayed in (2) the function is often called the logistic function form.

The Framingham investigators were concerned that the use of Fisher's discriminant analysis theory to estimate the regression coefficients of (1) and (2) might produce biased and inappropriate estimates since the formal assumption for that method is that the vector of risk factors ***X*** are multivariate normal. An assumption clearly not met since dichotomous variables such as cigarette smoking (CIG) and LVH were included in the models of Table 1. The decision was to shift from discriminant analysis and estimate the coefficients B conditional on the observed values of ***X***. This gave rise to **logistic regression** and a weighted least squares procedure was generated for the estimation (Walker and Duncan, 1976).

Table 2 contains prediction functions, called risk profile functions by the Framingham investigators at that time, which relate risk factors to the development of a first CVD event over an 8 year period (from Kannel, McGee and Gordon, 1976).

Table 2

Logistic Regression Functions for Relating Risk Factors to the Development of First Cardiovascular Disease (CVD) Event over an 8 Year Period (All risks factors are significant, $p < 0.05$)

	Men	Women
	Coefficients	
Constant	-19.7710	-16.4598
AGE	0.3743	0.2666
AGE^2	-0.0021	-0.0012
CHOL	0.0258	0.0161
SBP	0.0157	0.0144
CIG	0.5583	0.0395
LVH	1.0529	0.8745
GLUC	0.6020	0.6821
CH*AGE	-0.0004	-0.002

In the profile functions (or health risk appraisal functions) of Table 2 the hematocrit variable (HEM) of Table 1 has been dropped. The square of the age (AGE^2) and the interaction of age and total serum cholesterol (CH*AGE) have been added. The selection of risk factor variables of Table 2 became the standard set of variables for many Framingham predictive functions. Important references were published discussing the above mentioned issues, two of which are Halperin, Blackwelder and Verter (1971) and Gordon, Kannel and Halperin (1979).

3. THE USE OF MULTIPLE MEASUREMENTS ON A SUBJECT - METHOD OF POOLED REPEATED MEASURES (1968-1989)

As the Framingham study progressed a large amount of data accumulated from the biennial examinations (exams every two years). These data allowed the possibility to update risk factors for an individual and incorporate these into logistic regressions. The statistical method of pooled repeated measures was developed to achieve this (Cupples, D'Agostino, Anderson and Kannel, 1988). The method is basically a person examination approach and Table 3 indicates how it proceeds.

Table 3

Illustration of the Pooled Repeated Observation Method
Observations on Risk Factors taken every two years

PERSON EXAM APPROACH

	Exam *t*	*t*+1	*t*+2	sample
free of CHD	100	92	83	275 total
developed CHD	5	8	6	19 CHD
lost to follow up	3	1	2	

PERFOM ANALYSIS with 275 subjects and 19 events

At exam *t* there are 100 subjects free of CHD. Of these 5 develop CHD by exam *t*+1 and 3 are loss to follow-up. This leaves 92 subjects for exam *t*+1. Of these 8 and 1, respectively, develop CHD or are lost to follow-up by exam *t*+2, leaving 83 for exam *t*+2. Of these 83, 6 and 2, respectively, develop CHD or are loss to follow-up. Now summing the number of person exams we get 275 (= 100+92+83) and the number of events we obtain 19 (=5+8+6). In the context of the Framingham study, examinations would be separated by two years. The 275 and 19 are used as the data for say a logistic regression relating risk factors to the development of CHD within two years of the exam. Note the risk factors used in the analysis would be those obtained on the most recent examination.

The method of pooled repeated measures became a standard procedure in the Framingham study (Shurtleff (1974) and Cupples and D'Agostino (1987)). D'Agostino et al. (1990) showed that when used with the logistic regression, this method called in this context the pooled logistic regression, was asymptotically (i.e., for large samples) related to the Cox proportional hazard regression with time dependent covariates (D'Agostino, *et al.*, 1990).

Robert Abbott and Daniel McGee (1987) employed this method of pooled repeated measures using an eight year windows rather than two years as discussed above and generated health risk appraisal functions for a number of CVD events including myocardial infarction, CHD, CHD deaths, intermittent claudication, cerebrovascular accidents (strokes) and CVD.

These functions became extreme popular and heavy demands were made on the Framingham study to develop functions for specific purposes. For example, the Carter Center at Emory University asked the study to develop functions for mortality, while a number of drug companies asked for functions explicitly looking at blood pressure or cholesterol.

Also during this period Framingham publications appeared that clarified further the concept of these predictive functions (Gordon and Kannel (1982) and Kannel and McGee (1987)).

4. THE USE OF TIME TO EVENT ANALYSES (1987 - 1993)

4.1 Decision to Produce Authorized Functions

During the mid 1980s concern arose that too many health risk appraisal functions were developed and that the development was not undertaken in a systematic fashion with appropriate oversight control. Further, there was available newer statistical methods such as Cox proportional hazard regression that appeared ideally suited for these predictive functions but had not as yet been used. These methods were superior to the logistic regression in that they could take into account the time to event and also deal with drops outs and other forms of censoring. The study made the decision to generate "authorized" functions which would utilize the new methods.

4.2 Professional Functions and the American Heart Association

The first results of the above decision was to produce functions for professional use (Wolf *et al.*, (1991) and Anderson *et al.*, (1991)). The first function was a predictive function for stroke (Wolf *et al.*, (1991)). It employed a Cox proportional hazard model and can be used for estimating the probability of first stroke up to 14 years based on risk factors measured at baseline (i.e., time point 0). The mathematical model employed here is

$$S(t) = 1 - S_0(t)^{\exp(F)} \qquad (3)$$

where *S(t)* is the survival function at time *t* (that is, the probability of survival to at least time *t*) and

$$F = A + B_1 * X_1 + B_2 * X_2 + \dots + B_K * X_K \qquad (4)$$

The function $S_0(t)$ is called the underlying or average survival function and is the probability of survival at time *t* for a subject whose risk values at time 0 are equal to the mean values for all risk factors.

The second professional use function was for CHD. Unfortunately the proportionality assumption of the Cox regression did not hold for CHD and so a Weibull accelerated failure model incorporating a non-proportionality component was employed. The mathematical model for this function is given as

$$S(t) = EXP[-EXP\{(\ln(t) - F) / J)\}] \qquad (5)$$

where *S(t)* is the survival function at time *t* and *F* is a linear function such as (4)
and

$$\ln(J) = A + C * F \qquad (6)$$

The modelling of *J* as a function of the risk factors *X* accommodates non-proportionality. This CHD function can be used for estimating the probability of first CHD for 4 to 14 years following the measurement of the risk factors.

The CHD function was a single function employing 8 risk factors (sex, age, systolic blood pressure, total cholesterol, HDL cholesterol, cigarette smoking, presence of diabetes and LVH) and transformations of these. Sex specific stroke functions were produced, which included the above risk factors, except for total and HDL cholesterol, in addition to the existence of a previous CVD (but not a previous stroke), atrial fibrillation and the use of anti-hypertensive drugs.

The American Heart Association has distributed both the CHD and stroke functions to physicians to assess patients presenting to them with CVD risk factors. The presentation is devised in such a manner that 10 year probability estimates can be obtained directly once the risk factors are available. These can be compared to the average risk for someone of the same age and sex. These functions can also be used to estimate the effect of an intervention, because they can be use to estimate the change in probability if a risk factor is altered. For example, if a person quits smoking or if systolic blood pressure is reduced by, say, 20 mmHg.

4.3 Layman Functions for the American Heart Association

The American Heart Association also asked the Framingham study to produce health risk appraisal functions more suitable for the layman than the professional. This has been done for CVD and stroke. The CVD function goes under the label of **RISKO**.

5. SECOND EVENT MODELS (1994 - PRESENT)

Up to this point all the described health risk appraisal models have been for first event, first CVD, first CHD or first stroke. Current work has focused also on developing models for secondary events. By secondary event we mean an event that occurs in someone who already has had an event, such as the occurrence of a second myocardial infarction in a person who already has had a myocardial infarction. The Framingham work has focused on those who have survived the acute stage of the initial event. One set of sex specific Cox proportional hazard regression functions for predicting CHD or stroke in those with an initial CVD is as follows:

	Male	Female
	Coefficients	
ln(AGE)	1.006029	2.383863
ln(T-C/HDL-C)	0.957359	0.750673
CIG	0.0	0.764838
ln(SBP)	1.278776	1.157384
DIAB	0.229595	0.799036

the variable T-C/HDL-C is the ratio of total cholesterol to HDL cholesterol and the variable DIAB is the dichotomous variable for the presence or absence of diabetes. All other variables have been defined previously.

6. VALIDATION OF THE FRAMINGHAM FUNCTIONS

There is a substantial literature on the validation of the Framingham health risk appraisal functions, too broad to review here. However, it is worth mentioning that two recent articles reaffirm that they can be transported validly to other settings (Laurier, *et al.*, (1994) and Grover *et al*., (1995)).

7. SUMMARY AND ISSUES CURRENTLY UNDER INVESTIGATION

The Framingham study has been extremely successful in producing health risk appraisal functions. Much work continues. Some summary statements and issues still under investigation and development are as follows.

1. Health risk appraisal functions for CVD, CVD mortality, CHD, CHD mortality and stroke for follow-up periods extending from 2 to 14 years have been developed and are in use or ready for use.

2. New state of the art statistical methods have been incorporated into the development of these functions.

3. The stroke functions have incorporated hypertension medication. The incorporation of this into CHD functions is currently underway.

4. Approved professional models have been developed for CHD and stroke.

5. Layman model have also been developed for CVD and stroke.

6. Framingham models for secondary events have been recently developed. Preliminary functions will be published by the American College of Cardiology. Much further work is needed.

7. New variables such as triglycerides and fibrinogen are being incorporated into the models.

8. Functions are now being developed for different events such as cancer.

9. The Framingham health risk appraisal functions have proven to be valid in settings beyond Framingham.

REFERENCES

Abbott, R.D., and McGee, D. (1987). Section 37: the probability of developing certain cardiovascular disease in eight years at specific values of some characteristics, in Kannel, W.B., Wolf, P.A., and Garrison, R.J. (ed), *The Framingham Study, an Epidemiological Investigation of Cardiovascular Disease*, DHHS PHS NIH Pub 87-2284.

Anderson, K.M., Wilson, P.W.F., Odell, P.M., and Kannel, W.B. (1991a). An updated coronary risk profile: a statement for health professional. *Circulation*, 83, 356-362.

Anderson, K.M., Odell, P.M., Wilson, P.W.F., and Kannel, W.B. (1991b). Cardiovascular risk profiles. *American Heart Journal*, 121, 293-298.

Cupples, L.A., and D'Agostino R.B., (1987). Section 34: some risk factors related to the annual incidence of cardiovascular disease and death using repeated biennial measurements: Framingham heart study, 30-year follow-up, in Kannel, W.B., Wolf, P.A., and Garrison, R.J., (ed), *The Framingham Study, an Epidemiological Investigation of Cardiovascular Disease*, DHHS PHS NIH Pub 87-2703 (NTIS PB870177499), Washington, DC.

Cupples, L.A., D'Agostino, R.B., Anderson K., and Kannel W.B. (1980). Comparison of baseline and repeated measure covariate techniques in the Framingham heart study, *Statistics in Medicine*, 7, 205-218.

D'Agostino, Ralph B., and Kannel, W. B. (1990). Epidemiological background and design: the Framingham study, *Proceedings of the American Statistical Association Sesquicentennial Invited Papers Sessions - 1989 & 1988*, 707-718.

D'Agostino, R.B., Lee, M., Belanger, A., Cupples, L.A., Anderson, K. and Kannel, W.B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham heart study, *Statistics in Medicine*, 9,1501-1515.

D'Agostino, R.B., Wolf, P.A., Belanger, A.J., and Kannel, W.B. (1994). Stroke risk profile:adjustment for antihypertensive medication, *Stroke*, 25, 40-43.

Dawber, Thomas R., Kannel, W.B., and Lyell, L. (1963). An approach to longitudinal studies in a community: the Framingham study, *Annals of the New York Academy of Sciences,* 107, 539-556.

Dawber, Thomas R., Meadors, Gilcin F., and Moore, Felix E. (1951). Epidemiological approaches to heart disease: the Framingham study, *American Journal of Public Health*,41,279-286.

Gordon, T., and Kannel, W.B. (1982). Multiple risk functions for predicting coronary heart disease: the concepts, accuracy and application, *American Heart Journal*, 103, 1031-1039.

Gordon, T., Kannel, W.B., and Halperin, M. (1979). Predictability of coronary heart disease, *Journal of Chronic Diseases*, 32, 427-440.

Grover, S.A., Coupal, L., and Xiao-Ping, H. (1995). Identifying adults at increased risk of coronary disease, *Journal of the American Medical Association*, 274, 801-806.

Halperin, M., Blackwelder, W., and Verter, J. (1971). estimation of the multivariate risk function: a comparison of the discriminant function and maximum likelihood approach, *Journal of Chronic Diseases*, 24, 125-128.

Kannel, W.B., and McGee, D.L. (1987). Composite scoring - methods and predictive validity: insights form the Framingham study, *Health Services Research*, 22, 499-535.

Kannel, W.B., McGee, D., and Gordon, T. (1976). A general cardiovascular risk profile: the Framingham study,*American Journal of Cardiology*, 38, 46-51.

Laurier, D., Chau, N.P., Cazelles, B., Segond, P., and the PCV-METRA Group. (1994). Estimation of CHD risk in a French working population using a modified Framingham model. *Journal of Clinical Epidemiology*, 47, 1353-1364.

Shurtleff D, Section 30: some characteristics related to the incidence of cardiovascular disease and death: Framingham Study 18-year follow-up, in Kannel WB, Gordon T (eds), *The Framingham Study: an Epidemiological Investigation of Cardiovascular Disease*, US Government printing Office, DHEW publication (NIH) 74-599, 1974.

Truett, J., Cornfield, J., and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in the Framingham study. *Journal of Chronic Diseases*, 20, 511-524.

Walker, S.H., and Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables, *Biometrika*, 54, 167-179.

Wolf, P.A., D'Agostino, R.B., Belanger, A.J., and Kannel, W.B. (1991) Probability of stroke: a risk profile from the Framingham study, *Stroke*, 22, 312-318.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

INTERPRETING MULTIVARIATE TESTS

D.R. Thomas[1]

ABSTRACT

This paper discusses some of the procedures described in the literature for interpreting significant MANOVA tests. The focus is on measures that only require access to standard software packages such as SAS and SPSS. The measures recommended in this paper are the discriminant ratio coefficients introduced by Thomas (1992). They can be used to assess the relative importance of individual response variables to a significant multivariate test, and they can be used to help identify the underlying constructs associated with individual discriminant functions. Examples of their application are given.

KEY WORDS: Variable importance; MANOVA; Discriminant functions; Interpretation.

1. INTRODUCTION

1.1 Problem Description

This paper is concerned with the interpretation of significant multivariate tests arising from comparisons of group means, i.e., from the multivariate analysis of variance (MANOVA). Many statistics texts describe in detail the formulation and distributional properties of procedures for testing the equality of means of a set of response measures across two or more groups. These classical MANOVA tests, and corresponding p-values, are readily available to practitioners via statistical packages such as SAS and SPSS. Practitioners are also interested in procedures that will enable them to interpret those MANOVA tests that were declared significant. For example, they may wish to assess the relative importance of the response variables to a significant MANOVA, or they may wish to interpret the linear combinations of response variables, called discriminant functions, that are associated with each MANOVA test. Texts on multivariate analysis, particularly those written by statisticians, pay far less attention to this aspect of MANOVA. Whenever the topic is discussed, the recommendation is usually to examine individual discriminant function coefficients, or to examine the correlations between individual response variables and the discriminant functions. The continuing debate in the behavioural literature regarding methods for interpreting MANOVA tests has been reviewed by Thomas (1992). Some of the main issues in this debate will be summarized in this paper, and the shortcomings of existing approaches will be identified. Alternative measures of variable importance proposed by Thomas (1992) and Thomas and Zumbo (1995) will be described and illustrated on some real data sets. It will be shown that these measures, called discriminant ratio coefficients (DRC's), can be used to measure variable importance and can also be used to identify underlying constructs that may give rise to the linear discriminant functions. Recent unpublished work will be outlined in which multiple discriminant functions are rotated to maximise "simple structure" in the vectors of DRC's. An example will be presented to show that this technique can clarify the interpretation of the constructs that discriminate between the study groups.

1.2 Emphasis of the Paper

The emphasis throughout the paper will be on informal, or data analytic, procedures. In other words, there will be no discussion of the standard errors of the various coefficients that will be introduced. Also, it will be assumed that all sampled observations are independent and identically distributed within groups, i.e., there will be no discussion of weighted data or data drawn from clustered samples, though extensions to such complex samples are possible. With the exception of the experimental work on discriminant function rotation, the emphasis will be on methods of

[1] D. Roland Thomas, School of Business, Carleton University, Ottawa, Ontario, Canada, K1S 2T9.

interpretation that can be implemented by practitioners having access only to the standard MANOVA software provided by SAS, SPSS and similar statistical packages. Finally, the paper has been written to be accessible to practitioners who have no extensive knowledge of multivariate analysis. Some knowledge of basic matrix notation, plus a familiarity with the two-group t-test and the ANOVA F-test, should suffice.

2. MULTIVARIATE COMPARISONS: TWO GROUPS

2.1 Statistical Background

The necessary statistical background will be introduced in this section using a two-group example. Let $y_1, y_2, ..., y_p$ denote p response variables which can be represented as a p x 1 observation vector y. Consider two groups from which n_1 and n_2 observations are independently sampled, and let y_{jk} represent the k'th observation vector (k = 1, ..., n_j) from the j'th group ($j = 1,2$). Under the classical assumptions, the observations from the two groups will follow multivariate normal distributions with means μ_1 and μ_2, respectively, and common covariance matrix Σ. The multivariate group comparison then amounts to testing the equality of the mean vectors μ_1 and μ_2, i.e., to testing the multivariate null hypothesis

$$H_0: \mu_1 = \mu_2, \qquad (1)$$

where $\mu_j = (\mu_{1j}, \mu_{2j}, ..., \mu_{pj})', j = 1, 2,$ and the elements μ_{ij}, $i = 1, ..., p$ represent the means of the individual response variables, i.e., $\mu_{ij} = E(y_{ijk})$. Here E denotes expectation under the assumed multivariate normal model.

2.2 Example 1

This example is based on data from an investigation by Sénéchal, LeFevre, Hudson and Lawson (1995) of the relationship between children's verbal ability and their literacy environment. The subjects in the study were children aged from four to six, and for the purposes of this example they will be divided into two groups, one with high verbal skills (n_1=60), the other with low verbal skills (n_2= 59). It will be assumed that the two groups conform to the above statistical model, i.e., problems relating to misclassification will be ignored. In the original study, the data were analyzed using regression analysis so that grouping was not required. Four of the original study variables will be considered, each a measure of some aspect of a child's literacy environment. These are:

PRINTEXP	A measure of the primary adult caregiver's print exposure, measured by the proportion of the book titles on a given list that are recognized;
KNTTLSKB	A measure of the primary adult caregiver's knowledge of children's books, again measured as a proportion of titles recognized;
NUMKBKS	The number of children's books in the home;
NUMKREAD	The number of times the child is read to per week.

The analysis reported below is actually based on the square roots of these variables, since the distributions of the transformed variables were noticeably closer to normality than the those of the original variables. The names of the variables will not be changed.

2.3 Hotelling's T^2 Test

The appropriate test of hypothesis (1) is Hotelling's T^2. For the verbal ability data, the T^2 test results in a p-value less than .001, which, together with an effect size of 0.92 (see Stevens 1992, page 178), indicates a large difference between the vectors of group means. Separate univariate tests (Table 1) reveal significant group differences between the means of each of the individual response variables.

Table 1. Univariate Tests for the High and Low Verbal Ability Groups

Variable	t-statistic	d.f.	p-value
PRINTEXP	3.78	117	<.001
KNTTLSKB	3.96	117	<.001
NUMKBKS	3.72	117	<.001
NUMKREAD	3.22	117	.002

T^2 tests, and the measures used for interpreting them, can be best understood by posing the multivariate problem in univariate terms. For the verbal ability example, consider a new variable Z which is a linear combination of the four study variables, where for convenience the four variables PRINTEXP through NUMKREAD are denoted y_1 through y_4, respectively. Thus

$$Z = a_1 y_1 + a_2 y_2 + a_3 y_3 + a_4 y_4, \qquad (2)$$

where the a_i are fixed weights. The additional suffices representing groups and subjects within groups have been omitted from equation (2) for convenience. A specific weight vector $\boldsymbol{a}=(a_1,a_2,a_3a_4)'$ generates a value of *Z* for each subject in each of the two groups. Let *t(a)* denote the two-group *t*-statistic corresponding to this specific set of *a*'s. For example, when $\boldsymbol{a}$ = (.25, .25, .25, .25)', *t(a)* corresponds to a *t*-statistic computed using the average of each subject's scores on variables y_1 through y_4. Each value of $\boldsymbol{a}$ maps the multivariate data into a specific value of *t(a)*, or equivalently, of $t^2\,(a)$, since it is more convenient to work with positive quantities. It is natural to seek the weight vector a that yields the maximum value of $t^2\,(a)$, and the maximum value of $t^2\,(a)$ so obtained is Hotelling's T^2. The weights that yield the maximum are called discriminant coefficients, and the linear combination *Z* corresponding to the optimizing weights is usually referred to as Fisher's linear discriminant function. Because t-statistics are scale free, it is clear that T^2 is not affected if the vector $\boldsymbol{a}$ of discriminant weights is multiplied by a constant k. It is customary to chose *k* such that $\boldsymbol{a}^E=k\boldsymbol{a}$ satisfies $\boldsymbol{a}^{E\prime}\boldsymbol{S}_E\boldsymbol{a}^E = 1$, a process referred to as normalization (Thomas and Zumbo, 1995). Here $\boldsymbol{S}_E$ is the pooled within-groups estimate of the common covariance matrix Σ. As will be shown in the next section, the elements of the discriminant coefficient vector $\boldsymbol{a}^E$ and related coefficients are frequently used as aids to interpreting significant T^2 tests.

3. INTERPRETING A SIGNIFICANT T^2

3.1 Relative Importance of Variables

For many analysts, the first step in interpreting a significant multivariate test is to determine the relative contribution, or importance, of the individual variables. The term "relative importance", or simply "importance", of variables is often used in the applied literature, but is rarely defined with any precision as documented in the review by Kruskall and Majors (1989). A striking exception is Pratt's (1987) axiomatic development of a unique measure of variable importance for multiple regression. In the MANOVA context, an attempt at definition was made by Huberty and Wisenbaker (1992), whose "views" of variable importance included "contribution to discriminant function scores" and "contribution to grouping variable effects". Their recommended measures of variable importance will be discussed later in this section, as will competing measures that were introduced by Thomas (1992) and explored further by Thomas and Zumbo (1995).

3.2 Measuring Variable Importance

It is generally agreed that the univariate t-statistics of Table 1 do not provide useful multivariate measures of the importance of individual response variables to a significant multivariate test because they do not account for the correlations between the response variables. Instead, a number of multivariate methods have been proposed for measuring variable importance in MANOVA, several of which will be described in this section. The discussion will focus on the two-group T^2 case, while the multigroup case will be discussed in Section 4.

3.3 Discriminant Coefficients

It is natural to consider the discriminant coefficients as candidates when seeking multivariate measures of variable importance. For example, if coefficient a_2 in equation (2) is "large" in some sense, then it might be argued that the second variable is important because it is heavily weighted in the discriminant function that maximally discriminates between the groups. However, because different variables may be measured on different scales, discriminant coefficients must be interpreted with care. To make discriminant coefficients comparable, it is customary to standardize them, *e.g.*, by writing equation (2) in the form

$$Z=(a_1k_1)(y_1/k_1)+\ldots+(a_4k_4)(y_4/k_4), \qquad (3)$$

where the k_i are chosen to have the same scale as the *i*'th variable. The terms $a_i\;k_i$ appearing in equation (3) are referred to as standardized discriminant coefficients (SDC's), and denoted $b_i=a_i\;k_i$, $i=1,\;\ldots,\;p$. The choice of the scale quantities k_i has been the subject of considerable debate in the literature, reviewed recently by Thomas and Zumbo (1995). The most common choice is the standard deviation of y_i, i.e., $k_i=(S_{E\,ii})^{1/2}, i=1,\ldots,p,$ where $S_{E\,ii}$ denotes the *i*'th diagonal element of the within-groups sample covariance matrix $\boldsymbol{S}_E$ Thomas and Zumbo (1995) denoted the corresponding vector of SDC's $\boldsymbol{b}^{TT}$, where the first superscript refers to the standardization and the second subscript refers to the normalization. They also introduced an alternative set of SDC's, denoted $\boldsymbol{b}^{EE}$, for which normalization and standardization is based on "total" quantities instead of the within-groups quantities used to define $\boldsymbol{b}^{EE}$. These are preferable to $\boldsymbol{b}^{EE}$ (see Thomas and Zumbo, 1995), but must be calculated separately from the output provided by standard packages. The unstandardized and standardized forms are shown in Table 2 for the verbal ability data.

Table 2. Discriminant Coefficients for Comparing the High and Low Verbal Ability Groups

Variable	a^E	b^{EE}	b^{TT}
PRINTEXP	2.133	0.440	0.419
KNTTLSKB	2.365	0.352	0.337
NUMKBKS	1.070	0.342	0.325
NUMKREAD	0.410	0.287	0.269

Many authors implicitly assume that SDC's, suitably standardized, measure "contribution to discriminant function scores". Rencher and Scott (1990) explicitly recommended that the absolute values of the within-group standardized SDC's, b^{EE}, be used to assess the relative importance of variables to a two-group discrimination, i.e., to a significant T^2 test. However, Thomas and Zumbo (1995) argued that SDC's are not ideal measures of variable importance, irrespective of the standardization used. Different standardizations can lead to different importance orderings, an observation that sparked the earlier debate in the literature. For this and other reasons, Thomas and Zumbo (1995) recommended that SDC's be replaced as measures of importance by the discriminant ratio coefficients (DRC's) introduced by Thomas (1992).

3.4 Structure Coefficients

Most software packages routinely print the values of the within-group sample correlations between each response variable and each discriminant function. These correlations, referred to in the applied literature as structure coefficients (SC's), have also been proposed as measures of variable importance. However, their use in this regard has been discredited by the observation that, in the two-group case, the vector of structure coefficients, r^E, is proportional to the vector of t-statistics resulting from univariate tests of the individual response variables. Values of the SC's for the verbal ability example are shown in Table 3. It can be verified that these are proportional to the t-statistics shown in Table 1. Thus structure coefficients provide no multivariate information and are not useful measures of variable importance.

3.5 *F*-to-Remove Statistics

F-to-Remove statistics, denoted $F_{(i)}, i = 1, ..., p$, were originally recommended by Huberty (1984) as measures of variable importance in MANOVA. They were subsequently adopted by Huberty and Wisenbaker (1992) as the operationalization of their "contribution to grouping effects" view of variable importance. For the i'th response variable, *F(i)* can be obtained from an analysis of covariance of variable y_i, with all remaining p-1 response variables treated as covariates. $F_{(i)}$ is thus equivalent to Rao's (1973, page 551) test for the additional information contributed by y_i. $F_{(i)}$ is certainly a valid measure of variable importance. However, in multi-group cases, where there may be more than one significant discriminant function, $F_{(i)}$ provides only an overall measure of importance. It cannot describe the contribution of individual variables to each dimension of group difference.

3.6 Discriminant Ratio Coefficients

DRC's were originally proposed as measures of variable importance by Thomas (1992), based on a geometric interpretation of discriminant functions. Thomas also noted that DRC's are analogous to Pratt's (1987) axiomatically derived measures, an observation that further justifies their use as measures of importance. Thomas and Zumbo (1995) showed that DRC's are free of the deficiences listed above for SDC's, SC's and the F-to-Remove statistics, $F_{(i)}$. In the two-group case which features only a single discriminant function, the DRC's, denoted by d_i, i=1, ...,p, can be defined algebraically as

$$d_i = b_i^{EE} r_i^E = b_i^{TT} r_i^T, \qquad (4)$$

where r_i^E denotes a structure coefficient defined using "total" quantities in place of the within-groups quantities used to define r_i^E. Equation (4) states that DRC's can be defined using either total or within-groups quantities. Thus, using DRC's in place of SDC's as measures of importance sidesteps the debate regarding standardization of SDC's. Values of the DRC's for the verbal ability example are shown in Table 3, together with values of the *F(i)*'s, SDC's and SC's.

It can be seen from Table 3 that the DRC's sum to one, a property that provides an automatic scale for judging their relative magnitude, i.e., for deciding on the relative importance of the variables. It was shown by Thomas (1992) that each DRC can be interpreted in terms of lengths, *e.g.*, PRINTEXP accounts for 31.7% of the length of the discriminant function, when the latter is viewed as a vector in the space of the observations. It can be seen from Table 3 that for the verbal ability data, the F-to-remove indices, the within-groups SDC's and the DRC's all yield the same importance ordering, i.e., the variable that contributes most to the significant multivariate group discrimination signalled by the significant T^2 test is PRINTEXP, followed by variables KNTTLSKB and NUMKBKS, with NUMKREAD providing the smallest contribution. The importance orderings obtained using these different measures will not be the same in general.

Table 3. DRC's (and Other Measures) for the Verbal Ability Data

Variable	$F(i)$	b^{EE}	r^{E}	DRC ($b_i^E x r_i^E$)	Rank
PRINTEXP	3.69	.440	.719	.317	1
KNTTLSKB	1.99	.352	.753	.265	2
NUMKBKS	1.97	.342	.708	.242	2
NUMKREAD	1.52	.287	.613	.176	4
				1,000	

3.7 Further Notes on DRC's

Thomas and Zumbo (1995) discuss other aspects of the use of DRC's. First, though DRC's can be negative, negative DRC's of large magnitude signal the presence of highly correlated response variables. Large negative DRC's can therefore be removed either by using a "ridge" adjustment (see Thomas, 1992) or by dropping one or more of the highly correlated variables. Thus negativity of DRC's is no impediment to their use as measures of importance, an assertion also made by Pratt (1987) in the regression context. Second, DRC's can be used to identify suppressor variables, namely variables that make their contribution to the significant group discrimination through their relationship to the other response variables. A small DRC in conjunction with a relatively large value of $|b_i^{TT}|$ signals a suppressor variable.

4. MULTIVARIATE COMPARISON: MORE THAN TWO GROUPS

4.1 Statistical Background

Let $g > 2$ represent the number of groups from which n_j, $j = 1, \ldots, g$ observations have been independently sampled. The multigroup assumptions are a straightforward extension of those described in Section 2 for two groups, namely that observations in each group will follow multivariate normal distributions with mean vectors μ_j, $j=1, \ldots, g$, and common covariance matrix Σ. The multivariate null hypothesis in this case is given by H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$, where the elements of the mean vectors are defined as in the text following equation (1).

Multivariate tests of the multigroup null hypothesis can be derived by considering a linear combination Z of the p response variables, with p-vector of weights $\boldsymbol{a}$. For fixed $\boldsymbol{a}$, a univariate ANOVA on Z yields an F-statistic denoted $F(\boldsymbol{a})$, and as in the two-group case, multivariate procedures are obtained by maximizing $F(\boldsymbol{a})$ over all possible $\boldsymbol{a}$. Maximization leads to the canonical equation

$$\boldsymbol{Ha} = \lambda \boldsymbol{Wa}, \qquad (5)$$

where $\boldsymbol{H}$ and $\boldsymbol{W}$ are hypothesis and between-groups SSCP matrices, respectively. Equation (5) admits $g^* = min\ (p, g\text{-}1)$ solutions. When the given values λ_j, $j=1,\ldots,g^*$ are labelled in decreasing order, the weights $\mathbf{a}_1$ corresponding to λ_1 are the discriminant coefficients corresponding to the principal discriminant function, i.e., to the linear combination of variables that yields the maximum "univariate" F value.

Quotation marks are used here because the maximized F value does not follow the classical F-distribution. The second vector of weights $\boldsymbol{a}_2$, corresponding to the eigenvalue λ_2, yields the maximum value of $F(\boldsymbol{a})$ among all linear combination of response variables uncorrelated with the first, and so on. The eigenvalues are proportional to the g^* maximized values of $F(\boldsymbol{a})$. For the two-group case $g = 2$, equation (5) yields only one eigen-solution, with $\lambda_1 = 1 = T^2/(N-2)$, and $N = n_1 + n_2$. For the multigroup case $g > 2$, equation (5) admits two or more eigen-solutions. In this case, a number of multivariate test statistics are available, all of which are functions of the eigenvalues $\lambda_j 0$ (see Stevens 1992, page 226). A sequential test procedure is also available to determine how many of the possible g^* discriminant functions must be retained in order to fully describe the group differences (see, for example, Stevens 1992, page 275).

All measures of variable importance introduced in Section 3 for the two group case can applied in the multigroup case, with $g^* > 2$ discriminant functions. SDC's, SC's and corresponding DRC's can be defined separately for each discriminant function, the latter based on equation (4) as in the two-group case. SC's yield univariate information only and must therefore be discarded (Rencher, 1992). The previous criticisms levelled against SDC's still apply, leaving DRC's as the importance measures of choice whenever the contribution of individual variables to each significant discriminant function is of interest.

4.2 Example 2

This example is taken from an international marketing study carried out by Papadopoulos, Heslop and Bennett (1990). The subset of the data examined here relate to the perceptions of Russia and its people expressed by three groups of 150 respondents each from Canada, the United States and Australia. The six response variables that will be examined are:

ALIGN	Think we are (are not) aligned with Russia
IMMIG	Would (would not) welcome immigration from Russia
INDUS	Think Russia is (is not) industrialized
INVEST	Favour (do not favour) further investment in Russia
TIES	Should (should not) have closer ties with Russia
TRUST	Think Russians are (are not) trustworthy people

Responses are scored from one (negative) to seven (positive). A routine MANOVA test using the Wilks' L procedure (Stevens 1992, page 191) showed significant differences ($p < .001$) between the means of the three consumer groups, and a test of the residual effect attributable to the second discriminant function showed both dimensions of group difference to be important ($p = .008$). Thus the maximum number of discriminant functions ($g^* = 2$) will be retained in this example.

4.3 "Naming" the Two Dimensions of Significant Group Differences

DRC's can be used to determine the relative importance of the individual response variables to each dimension of group difference, i.e., to each discriminant function. For many analysts, the next step in the interpretation of a significant multivariate test is to identify the underlying constructs represented by the retained discriminant functions. Many authorities, *e.g.*, Huberty and Wisenbaker (1992), recommend using the structure coefficients for this purpose. However, Thomas (1992) suggested that DRC's should be used instead, and the DRC approach will be illustrated here using the consumer perception example. Table 4 displays SDC's and DRC's for both significant discriminant functions. It can be seen that variables ALIGN, IMMIG, INDUS and INVEST are the important contributors to the first discriminant function, while TRUST, ALIGN and IMMIG are, in order, the important variables contributing to the second discriminant function.

From the signs of the SDC's corresponding to the important variables, it can be seen from Table 4 that respondents will score highest on the first discriminant function if they:

think we ARE aligned with Russia
WOULD welcome immigration from Russia
think Russia IS NOT an industrialized country
FAVOUR further investment in Russia

Respondents will score highest on the second discriminant function if they:

think we ARE NOT aligned with Russia
WOULD welcome immigration from Russia
think Russians ARE trustworthy people

Thus the first discriminant function might be interpreted as a "political/economic" construct, while the second discriminant function might be interpreted as a "political/social" construct.

Table 4. SDC's and DRC's for the Consumer Perception Example

	First Disc. Function		**Second Disc. Function**	
Variable	b^{EE}	**DRC**	b^{EE}	**DRC**
ALIGN	.410	**.204**	-.479	**.160**
IMMIG	.556	**.368**	.450	**.278**
INDUS	-.465	**.209**	.211	.061
INVEST	.366	**.233**	-.096	-.024
TIES	.148	.062	-.330	-.040
TRUST	-.264	-.076	.803	**.564**

However, the presence of the variables ALIGN and IMMIG in both discriminant functions means that there is no clear separation between these two constructs. In factor analysis, a corresponding overlap between factors might be resolved by rotation of the factor loadings. Thomas (1995) investigated the merits of rotation of discriminant functions in MANOVA. He found that though rotation does destroy the maximization property of discriminant functions, it does preserve $\Sigma\lambda_j/(1+\lambda_j)$, i.e., it preserves the value of the Pillai-Bartlett criterion, one of the standard MANOVA test statistics. Thomas (1995) designed a rotation algorithm to maximize the "simplicity" of the DRC's, analogous to maximizing factor loading simplicity in factor analysis. Results for the consumer perception example are given in Table 5.

Table 5. DRC's After Discriminant Function Rotation for the Consumer Perception Example

	First Disc. Function		Second Disc. Function	
Variable	DRC	Sign (DRC)	DRC	Sign (DRC)
ALIGN	**.366**	+	-.002	-
IMMIG	.047	+	**.599**	+
INDUS	**.269**	-	.011	-
INVEST	**.145**	+	.064	+
TIES	.090	+	-.068	-
TRUST	.084	-	**.405**	+

Simple DRC structure has clearly been attained. The important contributors to the first rotated function are, in order of importance, ALIGN, INDUS and INVEST, and for the second function, IMMIG and TRUST. As a result of the rotation, the two discriminant functions can be associated with two substantively separate constructs, i.e., there is no overlap. The first construct can be interpreted as "political economic contact", the second as "social contact".

4.4 Displaying the Dimensions of Group Difference

Methodologists usually recommend that group averages of the discriminant function scores (group centroids) be plotted in the "discriminant plane", see for example Stevens (1992, page 277). A plot of the unrotated discriminant function centroids is shown in Figure 1, where the two discriminant functions are represented as orthoginal axes. This is not strictly correct. Figure 1 represents a two-dimensional subspace of the space of the response variables, so that the two discriminant axes are not orthogonal because the discriminant weight vectors a_1 and a_2 are not orthogonal. Nevertheless, this plotting scheme is frequently used and usually gives a reasonable interpretion of group differences. It can be seen that Canadian and US consumers rate Russia similarly on the "political/economic" dimension, in contrast to the Australians. On the "political/social" dimension, Canadian ratings are much higher than those of the Americans, while the Australians tend towards neutrality.

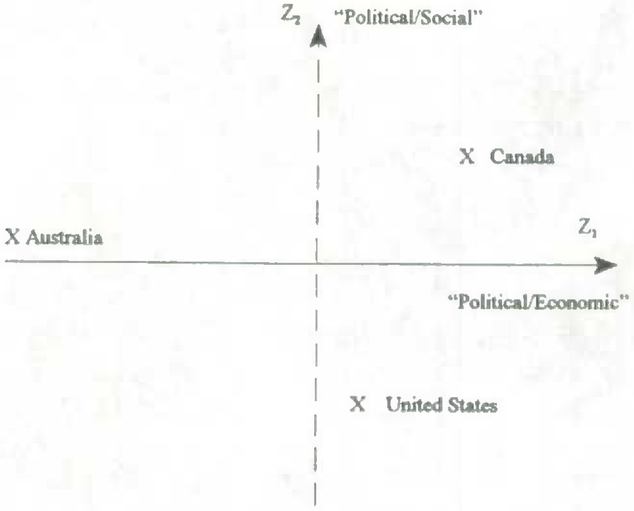


Figure 1. Group Centroids for the Consumer Perception Example in the Discriminant Plane

An accurate representation of group differences can be generated through a geometrical interpretation of MANOVA in the N-space of the observations, as described by Thomas (1992). Discriminant functions in the space of the observations provide orthogonal axes with respect to which the groups can be plotted as vectors. For the consumer perception example, this approach yields a two dimensional graph which confirms the interpretation of Figure 1. A similar approach provides a display of group differences with respect to the rotated discriminant functions, and again the conclusions for the consumer perception data are similar to the above. Further details are omitted for lack of space.

5. SUMMARY AND CONCLUSIONS

This paper has provided an overview of methods for interpreting significant T^2 and MANOVA tests. The focus has been on methods that only require access to standard MANOVA software. Following a review of the methods commonly described in the literature, the case for using discriminant ratio coefficients (DRC's) has been reviewed and illustrated. It has been shown that DRC's provide a basis for: (1) measuring the importance of response variables to each of the retained discriminant functions; (2) identifying possible underlying constructs associated with the discriminant functions; (3) defining a criterion for discriminant function rotation. A brief outline has also been presented of strategies for displaying group differences in relation to the discriminant functions. Sufficient detail has been provided to enable an analyst to explore the application of DRC's to the interpretation of any T^2 or MANOVA example. Analysts wishing to experiment with discriminant function rotation will find the required technical details in Thomas (1995).

6. REFERENCES

Huberty, C.J., and Wisenbaker, J.M. (1992). Variable importance in multivariate group comparisons, *Journal of Educational Statistics,* 17, 75-91.

Kruskall, W., and Majors, R. (1989). Concepts of relative importance in recent scientific literature, *The American Statistician*, 43, 2-6.

Papadopoulos, N., Heslop, L.A., and Bennett, D. (1993). National image correlates of product stereotypes: a study of attitudes towards East European countries, in F. van Raaj and G. Bamossy (eds.), *European Advances in Consumer Behaviour* (206-213), Amsterdam, The Netherlands: Association for Consumer Research.

Pratt, J.W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained, in T. Pukkila and S. Puntanen (eds.), *Proceedings of the Second International Conference in Statistics* (245-260), Tampere, Finland: University of Tampere.

Rao, C.R. (1973). *Linear Statistical Inference, 2nd. ed.* New York: Wiley.

Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components, *The American Statistician*, 46, 217- 225.

Rencher, A.C., and Scott, D.T. (1990). Assessing the contribution of individual variables following rejection of a multivariate hypothesis, *Communications in Statistics, Part B-Simulation and Computation*, 19, 535-553.

Sénéchal, M., LeFevre, J.-A., Hudson, E., and Lawson, E.P. (1995). Knowledge of picture books as a predictor of young children's vocabulary, Unpublished report, Department of Psychology, Carleton University, Ottawa, Canada.

Stevens, J. (1992). *Applied Multivariate Statistics for the Social Sciences*, New Jersey: Lawrence Erlbaum Associates.

Thomas, D.R. (1992). Interpreting discriminant functions: a data analytic approach, *Multivariate Behavioural Research,* 27, 335-362.

Thomas, D.R. (1995). Interpreting significant MANOVA tests: Discriminant ratio coefficients and discriminant function rotation, working paper, WPS 95-07, School of Business, Carleton University, Ottawa, Canada.

Thomas, D.R., and Zumbo, B.D. (1995). Using a measure of variable importance to investigate the standardization of discriminant coefficients, *Journal of Educational and Behavioural Statistics* (in press).

SESSION 3

Access and Control of Data

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

INFORMATIONAL PRIVACY AND DATA PROTECTION

D.C.G. Brown[1]

ABSTRACT

Higher volumes of information about people can now be processed at much higher speeds, which has lead to a growing concern among Canadians about their informational privacy. The federal government is the largest repository of personal information about Canadians. The *Privacy Act* sets out the rights of individuals to control their own personal information held by the federal government. While in some cases it may seem that privacy principles reduce efficiency, numerous government programs require personal information which will only be provided in an atmosphere of trust.

KEY WORDS: Privacy; Information collection; Privacy principles; *Privacy Act.*

It has become a truism to say that recent developments in information technology have had a profound effect on information processing and management, especially on the volume of information and the speed at which it can be processed.

Recent surveys have shown that as technology evolves, and as the public becomes more aware of the potential for the collection, retention and manipulation of personal information, Canadians are growing more concerned about the possible consequences for their informational privacy.

I am using the term "informational privacy" to describe the principle of control over one's personal information which has elsewhere been called "informational self-determination". Simply put it is the principle that individuals should know what information is being collected about them, by who and for what purpose and that (with few exceptions) they consent to the collection and use.

In addition to being a major collector, user and manager of information in general, the federal government is the largest repository of personal information about Canadians. Canadians provide a large volume of their personal information to the federal government for many purposes; for example, to apply for a variety of entitlements and benefits, to obtain a Social Insurance Number or a passport, in support of income tax payments, to obtain licenses, and, of course, in reply to the census.

The Canadian government first formally recognized the public's concern with the handling of their personal information by adding a section to the *Human Rights Act* which dealt with the handling of personal information by the federal government. In 1983 the general principles for the protection of personal information under the control of the federal government were expanded and clarified in the provisions of the *Privacy Act.* In addition, other pieces of legislation (such as the *Income Tax Act*, the *Statistics Act* and the *Unemployment Insurance Act*) contain provisions specifically governing the handling of personal information within particular programs or institutions.

Recent surveys have shown that the greatest concern individuals have about the collection and use of their personal information relates to the possible inaccuracy or misuse of information when making a decision concerning the subject individual. The collection and use of personal information for strictly statistical purposes is therefore not generally seen as a threat to individuals' privacy and is therefore not controlled as carefully.

The principles contained in the *Privacy Act* which relate most closely to the collection of statistical data would be:

- the requirement to inform individuals of the purpose for which information is being collected;

[1] David C.G. Brown, Executive Director, Information, Communications and Security Policy Division, Treasury Board Secretariat, Ottawa, Ontario, Canada, K1A 0R5.

- the prohibition on using personal information for a purpose other than the purpose for which it was collected;

- the requirement to provide individuals with access to their identifiable information upon request;

- the requirement to dispose of personal information in a manner consistent with its security classification; and

- the requirement to protect personal information from unauthorized disclosure.

I would like you to consider each of these principles in turn.

The requirement to inform individuals of the purpose for which their information is being collected. People fear a loss of control over their personal information if they do not believe that they know how their information will be used and how that use will affect them. This implies, of course, that those seeking information have to do their homework and define how the collected information will be used prior to beginning the collection. In addition to being informed as to the purpose of the collection, an individual who is asked to provide information for a government program must be informed as to the authority for requesting the information, whether the provision of the information is mandatory or voluntary and the possible consequences of refusing to provide the information.

The prohibition on using personal information for a purpose other than the purpose for which it was collected means that you cannot tell subjects that their information will only be used for statistical purposes (for example, to study the percentage of licensed pilots who also have a radio operator's license) and then decide to use the information for other purposes (possibly enforcement). The procedures which allow for new uses which are consistent with the original purpose include notification of the Privacy Commissioner, who determines whether it is appropriate to notify the subject individuals of the new use.

In order to satisfy the **requirement to provide individuals with access to their identifiable information upon request**, there has to be a method of tracing the distribution of identifiable information and of assigning accountability for that information so that it is clear who is responsible for responding to requests or complaints.

The **requirement to dispose of personal information in a manner consistent with its security classification** is an extension of the general principle of good information management which requires institutions to properly dispose of all government information holdings and supports the principle listed last here which is security for personal information.

The **requirement to protect personal information from unauthorized alteration or disclosure** means that you need appropriate security measures to ensure that your information is not vulnerable to access by people without authority and that the information is not improperly disclosed, altered or destroyed.

While most of the federal government programs which collect and process personal information have verification and enforcement mechanisms of some sort, they are still largely dependant on the willingness of the public to supply complete and accurate information about themselves. At present, members of the public are largely willing to entrust the federal government with their personal information with the expectation that their information will not be abused, or used for purposes they are not aware of. If, for some reason, the public trust were to decline, it could be expected that the quality of the information supplied voluntarily by the public would suffer. Federal programs would then need a great deal more resources for verification and enforcement and the predictive ability of the statistics drawn from these programs would decline sharply.

It is worth noting that some technological developments provide a means for the collection of some types of personal information without the knowledge of the individual. For example, the use of the "Number display" or "Name display" options offered by the telephone companies along with a "reverse directory" can allow businesses (or government institutions) to know who (or what household) has called. Among other uses, this information could be useful to a business in developing its marketing strategy.

The increasing number of "on-line" services may also result in the user being identified without their knowledge or in the tracking of their use of the system or services. These "transactional data" allow a data collector to develop a profile of an individual which may be useful for a variety of purposes. Federally, the *Privacy Act* prohibits such "data profiling" by government institutions without the knowledge of the subjects. The Treasury Board Manual on Privacy and Data Protection contains a chapter which specifically deals with data-matching by government institutions and describes the criteria for assessing and establishing a matching program.

As the data collection and manipulation capabilities of institutions have grown, there has been a growing temptation to attempt to amalgamate databases in order to achieve ever greater efficiency, however there is a delicate balance which needs to be maintained in order to benefit from the potential efficiencies while respecting the privacy principles I outlined earlier. Another factor to keep in mind is, of course, public perception. While there are some advocates for one centralized database of personal information (or even for a national identifier), such proposals are presently unacceptable to Canadians and seem to awaken the public's fear of "big brother".

Over the years Statistics Canada has demonstrated a sensitivity to the privacy concerns of Canadians which has served to reassure the Canadian public and keep unimpeded the flow of personal information needed by the department, not an easy feat in these days of growing public distrust of both government and technology, and I congratulate them on their achievements.

Those of us in the Information, Communications and Security Policy division of Treasury Board are happy to offer whatever assistance we can to any department which is facing the challenge of balancing the value of efficiency with the value of informational privacy, since we do not view these two values as necessarily being in competition. It is possible to achieve one without sacrificing the other, it may just take a little extra planning during the design of your information systems.

REFERENCES

Privacy Act, R.S., 1985, c. P-21.

Privacy Regulations, SOR/83-508.

Treasury Board Manual, Privacy and Data Protection, 1993.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

POPULATION REGISTERS AND PROTECTION OF PRIVACY: THE EXPERIENCE OF THE BALSAC REGISTER SINCE 1972

G. Bouchard[1]

ABSTRACT

The BALSAC register is a computerized database created from the linkage of parish records (births, marriages, deaths). The register covers the 19th and the 20th century and is now completed for the Saguenay and the Charlevoix regions. The work is continuing on the other regions of the province of Quebec, the whole of which is targeted (although at this scale, the data entry mostly involves marriage certificates only).

Owned by a consortium of Universities and managed by the Inter-University Institute for Population Research (IREP), BALSAC is only used for purposes of scientific research. Projects are being carried out within three programs relating to a) demographic and social structures, b) population genetics, c) cultural dynamics.

As is expected, the exploitation of the database is governed by a major set of controls and restrictions intended to secure an adequate protection of the confidentiality regarding data entry, storage, and utilization (such as: surveillance by the Quebec Commission of information access, public accountability, access permissions granted by external, independent bodies, users swearing in, contractual obligations, physical and computer restrictions and controls, etc).

KEY WORDS: Database; protection of confidentiality; population studies; IREP.

1. RIGHTS OF INDIVIDUALS AND NOMINATIVE DATA

The development and use of computerized nominative databases can to varying degrees cause problems with respect to protection of the confidentiality of the information and protection of the privacy of individuals. As used here, the term "confidentiality" refers to the character of some nominative data, the uncontrolled disclosure or dissemination of which (such as without the consent of the individuals concerned) may cause harm or infringe individual rights. By privacy, we mean the universe of personal or family information, over which each individual has a primary say. Thus, medical information is confidential insofar as its inappropriate disclosure may stand in the way of a promotion at work, tarnish a reputation, compromise a family situation, etc. While more inoffensive in appearance, personal information of an economic or cultural nature can have similar effects when disseminated in a specific context. In addition, linkage operations can change the nature of the data, making information confidential whereas it was originally not. A good example of this is the reconstituting of families from vital records: once linked, marriage and birth certificates can be used to calculate marriage and birth intervals and identify cases of premarital conception. In some circumstances, such linked records can also make it possible to detect so-called illegitimate births, identify persons related to subjects with hereditary diseases[2], etc.

Consequently, the management of a computerized nominative database should normally provide for a system or protocol for ensuring the protection of privacy. Such a protocol consists in a set of instructions or procedures setting out the conditions for collecting, storing, accessing, using and disseminating data. It must

[1] Gérard Bouchard, Director, IREP (Institut interuniversitaire de recherches sur les populations - inter-university institute for population research), 555 boul. de l'Université, Chicoutimi, Quebec, Canada, G7H 2B1.

[2] These examples may seem harmless, but our experience has shown us that for some of the individuals concerned, these are matters of some concern.

be constituted in such a way as to take into account, firstly, the legal and legislative framework of the community in question, and secondly, community moral and ethical standards regarding issues or matters not yet dealt with by lawmakers.

On the scale of Canada as a whole, the legal framework is determined in part by federal legislation, but also in part by provincial legislation, and hence there is considerable diversity. As to community sensibilities in ethical matters, they are also quite varied, as might be expected. Consequently, in the framework of the present document, it is completely impossible to propose a model that would apply throughout Canada, as the variety of contexts and situations calls for a variety of solutions. However, in order to provide food for thought and offer a sort of reference point among the various ones possible, it seemed useful to describe a specific approach, in this case the one followed by our Institute since it began its operations in 1972[3]. From it, everyone can draw at least an idea of the problems that are posed and the type of measures to take to overcome them.

2. THE BALSAC POPULATION REGISTER

The BALSAC register (which owes its name to the first letters of the names of several regions that it covers) is a particular type of computerized nominative database. By definition, a population register is a database that has the following characteristics:

2.1 the information that it contains (on occupations, places of residence, deaths, etc) must be explicitly attached to individuals;

2.2 such nominative data are connected through linkage programs, so as to bring together, at least theoretically, data relating to the same person;

2.3 to varying degrees depending on the records used, the data accumulated have a historical or retrospective dimension;

2.4 through genealogical filiation, the content of the register may be automatically scanned historically, over several generations;[4]

2.5 the register is supported by multiple-access softwares that can be used to "navigate" its various components (tables, fields, etc).

From a technical standpoint, such a structure can be created through the use of DBMS (database management systems) software. Concretely, the BALSAC register consists of a central register, which contains the description of all individuals involved (modules A, B and C of Figure 1; also Figure 2) and an indeterminate number of so-called sectoral registers containing specialized information of an economic, social, cultural or other nature. Implemented by using the INGRES management system, BALSAC now contains nearly 2 million baptism, marriage and death certificates from throughout the nineteenth and twentieth centuries. It currently covers the entire population of the Saguenay and Charlevoix regions (capture, linkage and validation completed), and it is now being extended over all regions of the province of Quebec (G. Bouchard, 1992) (Map 1).

This register, the construction of which began in 1972, is used in three research programs. The first is in the sphere of general social science (including geography and demography), the second is in the field of human genetics, while the third concerns cultural phenomena. In the first program, many surveys have been or are being carried out on subjects such as migratory movements, the decline in the birth rate, urban-rural relationships, the training of industrial manpower, family reproduction, socio-economic inequalities, etc. These surveys seek to reconstitute regional community dynamics and identify their disparities and discontinuities. In the second program, the work is divided between population genetics (kinship, founder effects, gene flow, etc) and genetic epidemiology. In the latter area, the work focuses on the dissemination and spatial distribution of genes, genealogical transmission chains, and the demographic, economic and social parameters of the risk of genetic disorders in populations. The objective pursued in this program is to determine the nature and tendency of the risk at the regional and interregional scale and thus help to prevent genetic diseases and diseases with a genetic component.

[3] IREP (Institut interuniversitaire de recherches sur les populations - inter-university institute for population research) is operated jointly by Université du Québec à Chicoutimi, Université Laval, McGill University, Université de Montréal, Concordia University and Université de Sherbrooke.

[4] On the above, see G. Bouchard *et al.*, 1985, 1989; G. Bouchard, 1988, 1992.

In the third research program, the work focuses on cultural dynamics (religious beliefs and behaviours, literacy, child-naming patterns, etc)[5].

3. PROBLEMS OF LAW AND ETHICS

Ethical and legal problems assume different forms. Our enumeration of these problems here is admittedly not exhaustive, and once again it reflects a specific approach, namely that of IREP, in a specific context. However, it may be seen that the issues identified are basically framed in fairly general terms.

3.1 Access to data and dissemination of information

The most basic concern relates to access to information and disclosure. There must be very strict controls to ensure physical protection of data (access to terminals, passwords, subregisters, etc). As a parallel measure, rules and procedures must be put in place for determining the accreditation of users and specifying the terms and conditions for data access and use. Most of the physical controls are standardized, but the rules for user accreditation may be quite variable, depending on the context.

3.2 Consent of individuals

As a rule, all nominative (and hence personal) information included in the database must have been authorized in advance by the persons concerned. But this principle does not apply when the data are public (for example, baptism, marriage and death certificates, land ownership data from registry offices, nineteenth-century Canadian census manuscripts, etc).

The problem of consent becomes more complex when (a) non-public nominative data are used without consent for research purposes; (b) public data are used for research purposes without having been created or authorized for those purposes; (c) by automatic linkage or another means, public data are converted to confidential data which are then used without the consent of the persons concerned; (d) two nominative databases or registers are linked in order to produce a third register, of a different nature.

3.3 Inference and interference

The construction of genealogies may make it possible to reconstitute harmful gene transmission circuits or estimate (by "genealogical inference") the probability that such genes will be disseminated among offspring and families. The use of this knowledge may lead to serious breaches of privacy, where the researcher, acting in the interests of prevention, could be tempted to interfere.

3.4 Wrongful uses of registers

A register constructed for research purposes may later lend itself to use by commercial or industrial firms for other purposes (e.g. personnel selection), or by government departments. Such possibilities give rise to the fear that data will be re-used or even retrieved or "perverted" in unforeseen ways, possibly to the detriment of human rights. One need only think of the potential for linking a register with a set of prospective clients of insurance companies.

3.5 The citizen's right to participate in decisions

Thus the use of a population register, like any nominative database, lends itself to forms of use which were not necessarily foreseen at the outset and which can violate society's ethical principles. Operating within the closed confines of their universities, researchers are tempted to make, without public involvement or consultation, scientific decisions that are in fact social choices of the highest order, for which they have no mandate whatsoever. This poses the problem of collective responsibility for formulating research goals.

3.6 Long-term guarantees

The construction of a population register is the work of a generation of researchers endeavouring to achieve clearly defined scientific objectives. Once those objectives are achieved, the original team is dissolved, but the scientific infrastructure that has been set up remains. Thus it is not impossible that under the certain circumstances, the infrastructure will continue its "career" in a different, less secure environment, at the mercy of initiatives prejudicial to the rights of the individuals concerned.

All nominative databases are of a nature to give rise to one or another if not all of the problems described above. For this reason, it is important for the scientific community to help build general awareness of the issues and contribute to a process of reflection which in any

[5] For all the preceding, see IREP Documents, Nos. I-C-126, I-C-134, I-C-138. Also see the annual reports published by the Institute.

event has been under way for several years[6].

4. THE IREP PROTOCOL

With a view to providing an adequate solution to each of the problems identified, IREP and Université du Québec à Chicoutimi have established a relatively complex protocol that now governs the management of the BALSAC register. Its main features are as follows. It should be noted at the outset that four universities (Université du Québec à Chicoutimi, Université Laval, McGill University, Université de Montréal – hence four public bodies) own the population register and are responsible for its management. Back in 1977, with the help of legal experts, IREP[7] and Université du Québec à Chicoutimi instituted a protocol governing data access, storage and use. This first protocol was made up of instructions and procedures providing for physical or technical measures (passwords, access to premises, system features providing for restricted access and compartmentation of computerized files, etc), as well as contractual obligations imposed on persons working for IREP (e.g. the rule of anonymity in published or disseminated findings or data)[8]. On the latter point in particular, the protocol stipulated that research staff and register users be sworn to secrecy and that various types of contracts (for hiring of assistants, construction of subregisters, loaning of data, etc) provide for various penalties in the event that the rules were breached. It also gave a ten-member committee, independent of IREP, the responsibility for managing access to the data and generally applying the protocol. Lastly, it prohibited the use of the register for commercial purposes.

In 1980-82, this operating framework underwent an in-depth reevaluation and revision, this task having been assigned to a team of legal scholars led by Professor Jean Goulet of the Faculty of Law of Université Laval. The new protocol that resulted from these efforts retained the previous provisions and expanded on them. It also ensured that the new regulations conformed to the quite recent federal and Quebec legislation[9]. As a result of these changes, the management of the register with respect to matters of ethics and law was under the purview of three bodies: a university-wide ethics committee operating at arm's length from IREP, the board of directors of Université du Québec à Chicoutimi (through the office of the secretary general), and the Quebec government's Commission d'accès à l'information (J. Goulet *et al.*, 1983). From that point on, any request for access to data, the creation of subregisters or the launching of research projects had to be approved by those three bodies[10]. In turn, this second protocol was amended several times during the 1980s (IREP, 1989).

With respect to the matter of consent, the protocol obliges researchers to obtain the consent of all persons providing information. Where it is impossible to satisfy this obligation (for example in the case of a body of data concerning thousands of persons, some of whom are deceased), IREP is authorized to take exceptional measures contemplated for this purpose by the legislation of Quebec and Canada[11].

As regards more specifically the use of the register in the field of human genetics, in particular for the purposes of genetic epidemiology, it seemed justified to develop a subset of rules appropriate to the type of problems posed in this sphere. The basic question may be posed as follows: when and insofar as the database, through genealogical analysis or otherwise, may yield information of a quasi-medical nature on individuals (e.g. information regarding the risk of carrying a given harmful gene), how and under what conditions may this information be used? The policy developed is based on the following guidelines. First, there can be no question of tracking down carriers of mutant genes by using the

[6] Among the contributions worthy of note are those of our colleague David Flaherty of Western University (London, Ontario). As regards IREP, see in particular J. Goulet (1992), C. Laberge, B.-M. Knoppers (1992), B.-M. Knoppers, C. Laberge, Loïc Cadiet (1992), G. Bouchard (1993).

[7] IREP was then known as SOREP (Société de recherches sur les populations -- society for population research).

[8] For the system of physical protection, see IREP Document I-C-148.

[9] For example, the 1982 Quebec Act respecting access to documents held by public bodies and the protection of personal information (RSQ, c A-2.1).

[10] A modus operandi was established with the Commission d'accès in order to simplify the authorization process.

[11] For Quebec, see section 19 of the Act respecting health services and social services. Also see section 59 (subsection 5) and 125 of the Act respecting access to documents held by public bodies and the protection of personal information (RSQ, c A-21).

register in operations focusing on offspring identified as being at risk on the basis of genealogical inference. The researcher would then be breaching the basic rules of law and ethics by intruding in the lives of individuals and families without having been expressly requested to do so by the persons involved. Second, the dissemination and use of information regarding risk or any other personal information of a medical or quasi-medical nature drawn from the register must be handled by authorized medical authorities, in the framework of genetic counselling consultations between health professionals and individuals requesting such consultations.

Lastly, a request for access to the register from an individual for any purpose other than scientific research is acceptable only if the information sought concerns only the individual making the request and on condition that the information has no medical or genetic connotation. Under the existing protocol, such a request must then be directed to the office of the secretary general of Université du Québec à Chicoutimi for approval. However, a request of this nature is not approved if it involves information of a medical or genetic nature. In the latter event, the individual making the request is referred to a competent medical unit.

Theoretically, the BALSAC register could make it possible to assemble all the information concerning a given person. But in practice, there is no such thing as an "individual file." In computer terms, such a file is only virtual, since the information concerning a given individual is distributed among a number of subregisters (or "tables") that have deliberately been structured in such a way that it is impossible for the user to interrelate them all. This operation, which could be performed only by the managers of the register, would require complex computer manipulations and would have to have first received outside authorizations.

For the past two years, IREP has been involved in a thorough revision of the confidentiality protocol, which is now undergoing its third redesign. The operation, once again carried out under the supervision of legal scholar Jean Goulet, is to be completed during 1996. The new protocol takes up the principles and the main rules of the old one, adapting them and makes various additions and refinements to them. The guidelines that underlie the protocol are set out in Table 1. The structure of the system is illustrated in Figure 3.

A final point deserves attention, namely the matter of collective rights. Here we are referring to researchers' obligation to protect the image of the groups which they are studying. This question is especially delicate when it comes to research on genetic diseases. In the absence of appropriate precautions surrounding the terms and conditions for dissemination of findings, it is dangerous to generate negative stereotypes that long afterward remain associated with the population groups in question. An awareness of this problem led IREP to take the initiative of drafting a sort of code of ethics which has been proposed to the members of the scientific community and various groups involved in the field of hereditary diseases; it has also been brought to the attention of media professionals[12].

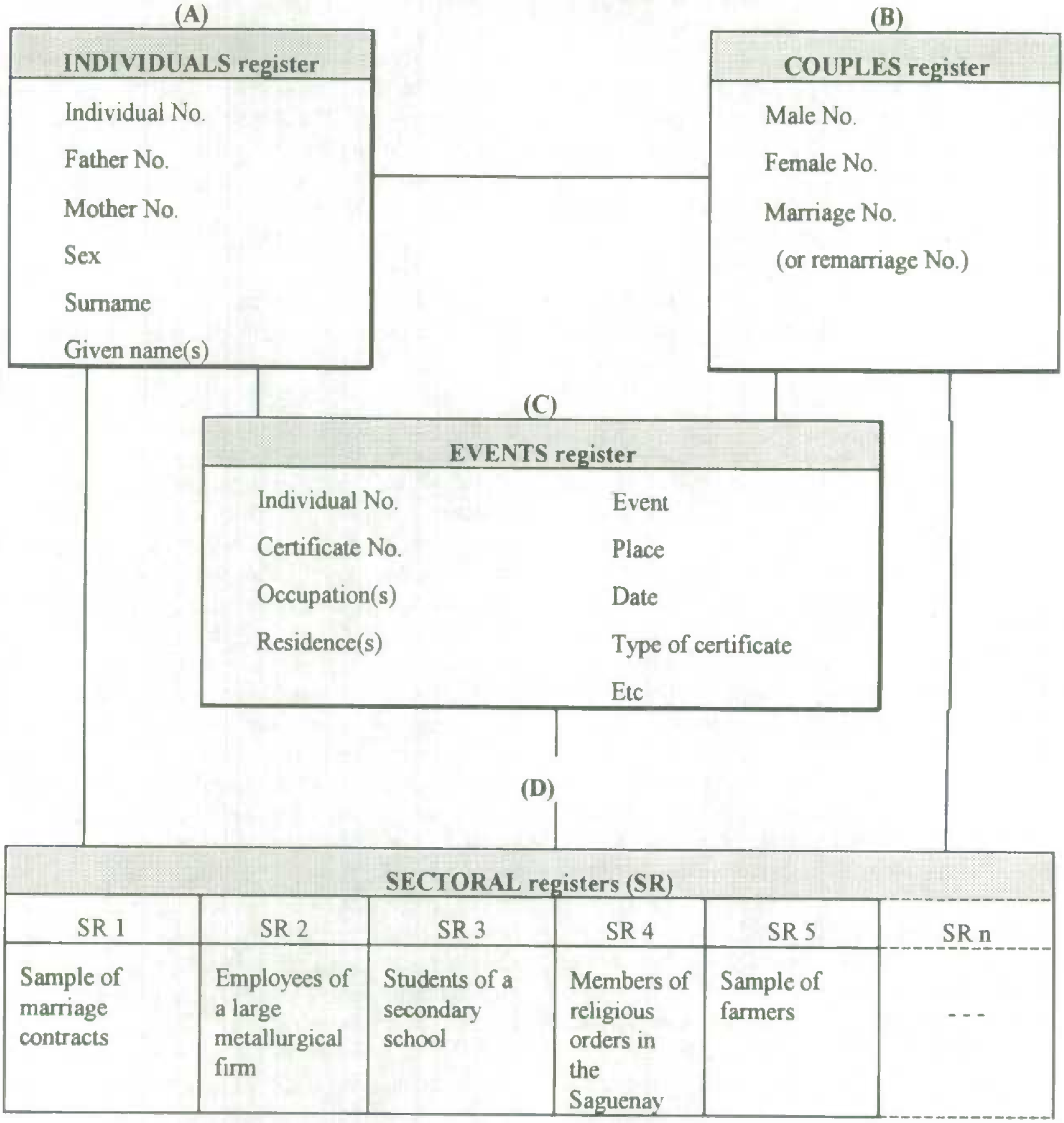
5. CONCLUSION

As a general rule the IREP protocol favours transparency in research operations and decisions. Regularly, IREP announces important upcoming developments via the media. Regularly too, the policy adopted with respect to confidentiality is the subject of presentations and discussions in symposiums and seminars which are open to the public and which bring together experts in law, ethics and various disciplines. Thus the multidimensional structure for decision-making and consultation, grounded in the university, provides protection against the possibilities of commercialization or "perversion" of the register. Supported by public institutions, the register is reasonably sheltered from unforeseen developments that could endanger its long-term future. We also believe that the protocol provides adequate protection on each of the six points enumerated above (Part 3). On that score, it is worth noting that over a period of twenty years, the operation of the register has not resulted in any incident involving harm to individuals or breach of privacy, nor has it given rise to any complaints by aggrieved individuals. However, it is important to recall that the situations created by nominative databases are constantly evolving, owing to ongoing technological change, changes to the legal and legislative framework, and changes in collective sensibilities in favour of respect for human rights. It goes without saying that the scientific community must wholeheartedly support the latter trend and show a willingness to continually alter its scientific practices in order to make them conform to the fundamental principles of collective morality. For all these reasons, it must be clearly recognized that protocols on the protection of privacy are always provisional and must be regularly subject to review, in a spirit of inquiry that is as open as possible.

[12] On this subject see G. Bouchard (1994) and IREP Document No. III-C-94.

Figure 1

Structure and Content of the BALSAC Population Register

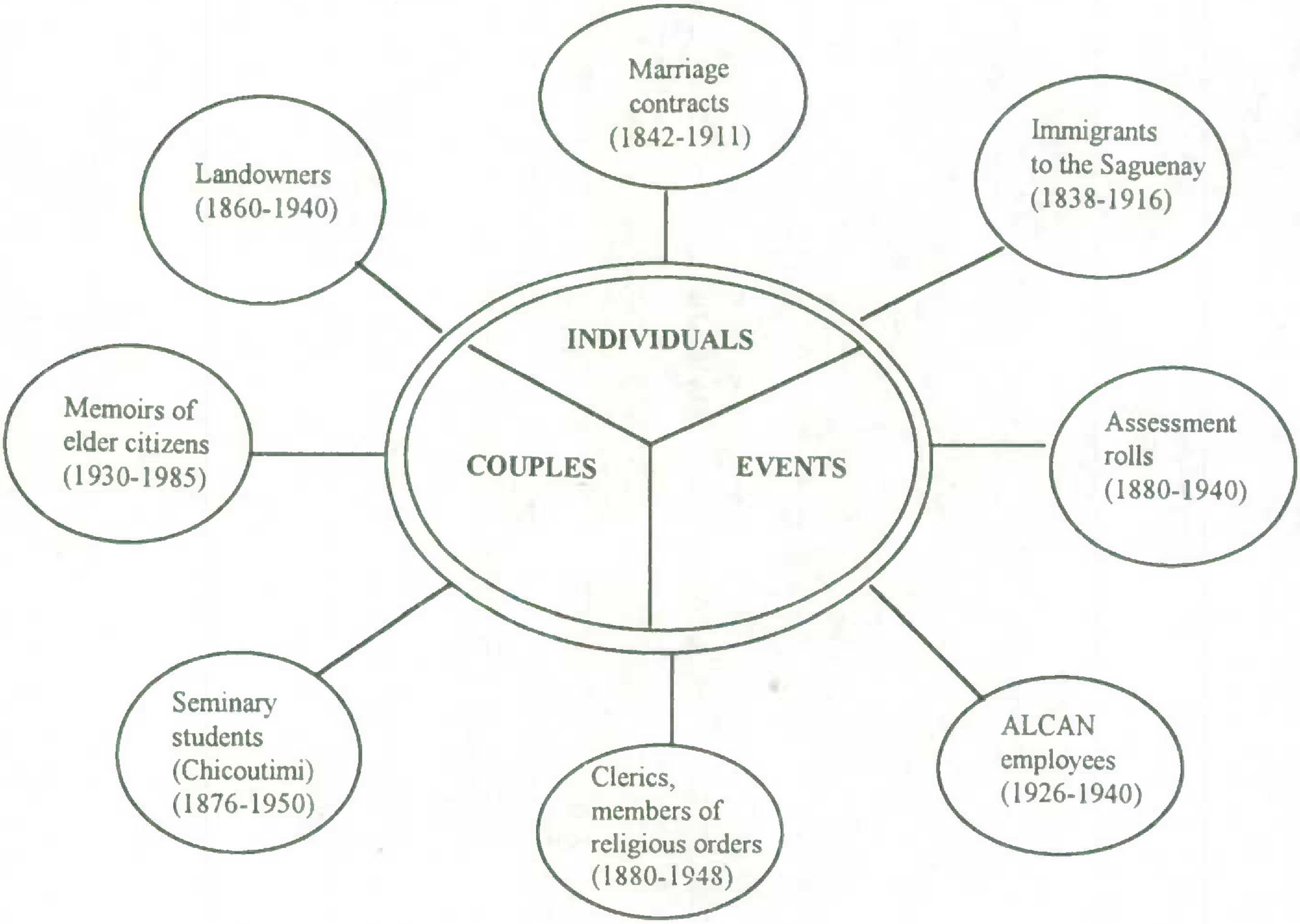


(IREP)

Figure 2

STRUCTURE AND CONTENT OF THE BALSAC POPULATION REGISTER

(Central Register. Examples of Sectoral Registers)


Marriage contracts (1842-1911)
Immigrants to the Saguenay (1838-1916)
Landowners (1860-1940)
INDIVIDUALS
Memoirs of elder citizens (1930-1985)
COUPLES
EVENTS
Assessment rolls (1880-1940)
Seminary students (Chicoutimi) (1876-1950)
Clerics, members of religious orders (1880-1948)
ALCAN employees (1926-1940)

(IREP)

Map 1

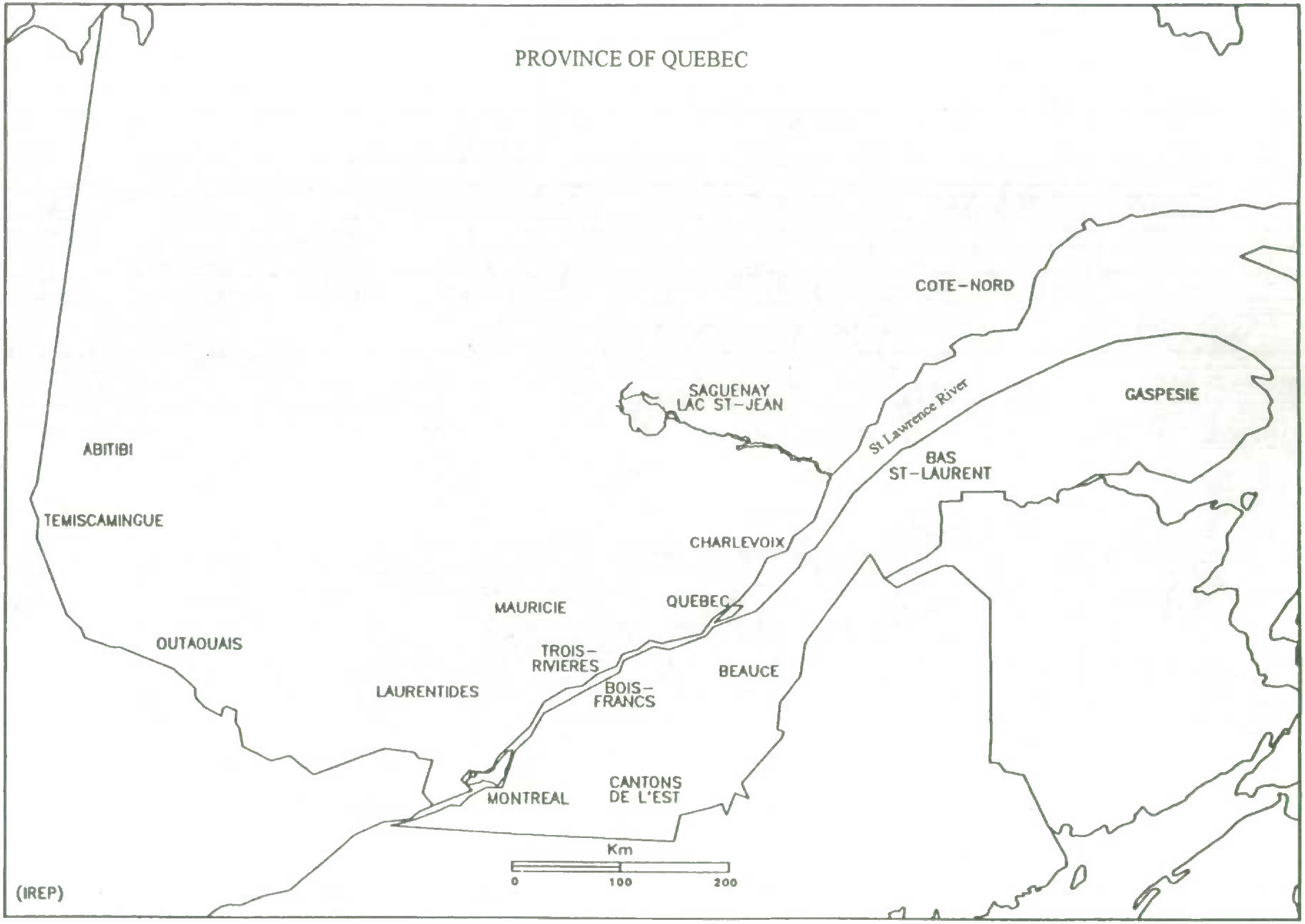

PROVINCE OF QUEBEC
COTE-NORD
SAGUENAY
LAC ST-JEAN
St Lawrence River
GASPESIE
BAS
ST-LAURENT
ABITIBI
TEMISCAMINGUE
CHARLEVOIX
QUEBEC
MAURICIE
OUTAOUAIS
TROIS-
RIVIERES
BEAUCE
BOIS-
FRANCS
LAURENTIDES
CANTONS
DE L'EST
MONTREAL
Km
0
100
200
(IREP)

Table 1

CONFIDENTIAL PROTECTION SYSTEM GOVERNING THE USE OF THE BALSAC REGISTER (MAIN FEATURES)

1 - CONTENT OF REGISTER: PUBLIC DATA

2 - OWNERSHIP OF REGISTER: OWNED BY FOUR UNIVERSITIES

3 - NO INTRUSION ON PRIVACY

4 - NO USE OF THE REGISTER TO OBTAIN PROFIT FOR EITHER ITS OWNERS OR MANAGERS (SCIENTIFIC RESEARCH ONLY, APPROVED BY AN ETHICS COMMITTEE)

5 - EXTERNAL MECHANISMS FOR AUTHORIZING ACCESS

- UNIVERSITY-WIDE ETHICS COMMITTEE
- SENIOR MANAGEMENT OF UNIVERSITÉ DU QUÉBEC À CHICOUTIMI
- COMMISSION D'ACCÈS À L'INFORMATION DU QUÉBEC

6 - REGISTER DOES NOT RETAIN MEDICAL DATA[a]

7 - RESTRICTED-ACCESS SOFTWARE (DATA COMPARTMENTATION MECHANISM)

8 - IN PRACTICE, THERE IS NO SUCH THING AS AN "INDIVIDUAL FILE" (FRAGMENTATION OF DATA)

9 - DISSEMINATION OF RESEARCH FINDINGS: RULE OF ANONYMITY

10 - CONTRACTS FOR SWEARING-IN OF USERS AND EMPLOYEES, ACCOMPANIED BY PENALTIES

11 - CONSTANT MONITORING OF OPERATIONS BY A CONTROL COMMITTEE (25 - 30 MEETINGS PER YEAR)

12 - CONCERN FOR PROTECTING BOTH THE IMAGE OF GROUPS COVERED AND INDIVIDUAL REPUTATIONS

[a] See IREP document No. I-C-153.

Figure 3

Structure of Confidentiality Protocol
Types of protection controls governing access to and use of the BALSAC register network

Legal and legislative	Institutional (internal and external)	Personal, contractual	Technical, computer	Physical
• Civil Code of the province of Quebec • Charter of Rights and Freedoms • 1982 Act respecting access[a]	• Control committee • University-wide ethics committee • Board of directors of the Université du Québec à Chicoutimi • Commission d'accès à l'information du Québec	• Swearing in of users • Penalties • Banalized results • Contractual obligations	• Passwords • Restricted access • Compartmentation of subregisters • Dispersion of individual data	Controlled access: • premises • terminals • computer • disks • tapes • lists • etc.

(a) Act respecting access to documents held by public bodies and the protection of personal information (RSQ, c A-2.1). The register is jointly owned by four universities, and as the latter have the status of public bodies, its management comes under this Act.

REFERENCES

Bouchard, G. (1988). Les fichiers-réseaux de population: Un retour à l'individualité, *Social History/Histoire sociale*, XXI, 42, 287-294.

Bouchard, G. (1992). Current issues and new prospects for computerized record linkage in the province of Québec, *Historical Methods*, 25, 2, 67-73.

Bouchard, G. (1992). *Le Centre interuniversitaire SOREP et le fichier BALSAC, État présent et planification des travaux.*

Bouchard, G. (1993). Retracer les gènes dans la population: une infrastructure de recherche pour le 21e siècle, *Transactions of the Royal Society of Canada*, sixth series, IV, 13-24.

Bouchard, G. (1994). Les problèmes de droit et d'éthique reliés à l'exploitation d'un fichier de population à des fins génétiques, Marcel J. Mélançon (dir.), Bioéthique et génétique, Une réflexion collective,. Chicoutimi, Éditions JCL, 33-42.

Bouchard, G., Roy R., and Casgrain B. (1985). *Reconstitution automatique des familles, Le système SOREP*, Dossier no. 2, Université du Québec à Chicoutimi, 2, 745.

Bouchard, G., Roy R., Casgrain B., and Hubert M. (1989). Fichier de population et structures de gestion de base de données: le fichier-réseau BALSAC et le système INGRES/INGRID, *Histoire & Mesure*, IV, 1/2, 39-57.

Goulet, J. (1992). La législation sur la protection de la vie privée: les principes fondamentaux des lois de première génération, *Les archives non textuelles: réflexions théoriques et expériences pratiques*, Proceedings of the colloquium organized jointly by the archives division and the records science program of Université Laval, November 20, 1991, Québec: Université Laval, 103-119.

Goulet J., Gagné M., and Girard D. (1983). Règles de droit et confidentialité, Dossier no. 1, 175.

IREP (1989). *Règlement concernant la confidentialité des données contenues dans le fichier BALSAC*, March, 40.

Knoppers, B.-M., Laberge, C., and Cadiet, L. (dirs.) (1992). *La génétique de l'information à l'informatisation*, Proceedings of the colloquium held at the faculty of law of Université de Montréal, organized by the C.R.D.P., Université de Montréal and the C.R.J.O., Université de Rennes, Paris: Litec, 387.

Laberge, C., and Knoppers B.-M. (dirs.) (1992). *Registres et fichiers génétiques: enjeux scientifiques et normatifs*, Montréal: Association canadienne-française pour l'avancement des sciences (ACFAS), Collection Les cahiers scientifiques, 77, 178.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

LEGAL/POLICY ASPECTS OF CONFIDENTIALITY AND PRIVACY

L. Desramaux[1]

ABSTRACT

The presentation will describe the legislative/policy framework which governs Statistics Canada's activities with respect to collection, use and disclosure of information with particular emphasis on personal information. It will focus on the legal basis for this framework as found in the *Statistics Act*, the *Privacy Act* and the *Access to Information Act*. It will also describe the policies that the Agency has developed to support the requirements of legislation such as the Record Linkage Policy, the Policy on Informing Survey Respondents and the Microdata Release Policy.

KEY WORDS: *Statistics Act; Privacy Act; Access to Information Act* ; Policy on Informing Survey Respondents; Microdata Release Policy.

1. INTRODUCTION

The purpose of this presentation is to provide an overview of the legislative/policy framework that Statistics Canada has adopted to help ensure that the Agency properly discharges its responsibilities with respect to data collection, protection and disclosure.

Statistics Canada has two fundamental responsibilities. One is to inform the public, to the best of its ability, and the other is to ensure the confidentiality of individual information provided to it.

These two responsibilities are intimately linked. To respond to information requests, what is required is a stable information source, which can only be ensured if the commitment to respondents to protect the information they provide is respected.

Not only are these responsibilities inextricable; they can sometimes conflict. Society needs more information to understand increasingly complex problems, and to accomplish this, it needs increasingly detailed information. On the other hand, it should be noted that public attitudes are changing with respect to information collection. Individuals appear to be more concerned about the accumulation of information collected in their regard, and about the use made of this information. They are also concerned about the power of technology, which among other things makes it possible to link information from a number of sources, thus creating massive data banks on individuals. Additionally, they are increasingly aware of the rights vested in them by federal and provincial legislation on privacy.

To be responsibly and consistently responsive to the public's concerns about increased accumulation of information, particularly personal information, and to legitimate pressures to provide more detailed information, Statistics Canada has put in place a legislative/policy framework that focuses exclusively on questions of confidentiality, privacy and security.

The legal basis for this framework is the *Statistics Act*, the *Privacy Act* and the *Access to Information Act*. It is supplemented by a number of well-documented policies and procedures which support compliance with the legal requirements.

2. THE STATISTICS ACT

The cornerstone of Statistics Canada's legislative/policy framework is the *Statistics Act*. This act was first proclaimed in 1918. It was extensively revised over the years, the last major revision occurring in 1971. It sets out Statistics Canada's mandate which is to:

- collect, compile and publish information;

[1] Louise Desramaux, Data Access and Control Services, Statistics Canada, Ontario, Ottawa, Canada, K1A OT6.

- take the census of population and agriculture;
- collaborate with departments of government;
- promote and develop integrated social and economic statistics;
- promote avoidance of duplication with other government departments.

To effectively carry out such a mandate, there must be three legislative requirements: the authority to collect information, the obligation on the part of the respondent to provide the information, and the protection of the confidentiality of the information once provided.

2.1 Authority to Collect Information

The *Statistics Act* gives the Chief Statistician broad powers of collection, covering a wide array of subjects. It allows for collection directly from respondents as well as for accessing administrative records held by departments of government, federal, provincial and municipal as well as records held by any corporation.

2.2 Obligation to Respond

Since it was first proclaimed in 1918, the *Statistics Act* required that response to any Statistics Canada surveys be mandatory. During the 1970s, information required by governments and researchers to address emerging social issues necessitated the collection of information that respondents considered to be intrusive. Some of these surveys were becoming more complex and at times required lengthy responses. The Agency was fully aware that the more sensitive nature of some of its statistical enquiries called for surveys that allowed for a voluntary response. Consequently, in 1981, the *Statistics Act* was amended to allow the Minister, by order, to authorize the taking of a survey to which response is voluntary. The Minister, however, according to these provisions cannot make the Census of population or the Census of agriculture a voluntary survey.

2.3 Secrecy

As a counterbalance to the extensive collection powers and the obligation in some instances on the part of respondents to provide information, the *Statistics Act* contains very stringent confidentiality provisions. There are two parts to the section of the Act concerned with secrecy. The first part deals with access to identifiable returns collected under the *Statistics Act*. No one can have access to individual information unless he/she has been sworn in under the legislation.

The second part deals with what information can be disclosed. The wording is rather awkward but what is important to retain is that it is very constraining. In a nutshell, the requirements are that when information is released, it must not be possible to relate anything on a return to an individual respondent. They have led to the development of a number of disclosure control methods to deal with the practical aspects of ensuring confidentiality. Release of individual information has been permissible only since the *Statistics Act* was revised in 1971. Prior to that, Statistics Canada could only disseminate aggregate data. There had been exceptions to this, even as far back as 1918. For example, there was some discretion then in releasing individual transportation returns. Over the years exceptions were broadened to include returns of carriers and public utilities, returns of institutions such as hospitals, mental institutions, libraries, educational institutions as long as individual patients, inmates or other persons in the care of such institutions could not be identified. There was also some discretion in releasing returns of any respondent, be they a business or an individual with the written authorization of the respondent and in releasing lists of businesses that would contain the name and address of businesses, the type of business, the size of the business as expressed by the range of number of employees. This discretion can only be exercised by the Chief Statistician, requires him to issue an order and is governed by an internal policy.

While these provisions allow some flexibility in releasing information that would otherwise be suppressed, by and large, the confidentiality provisions are, as previously mentioned, very strict. There is mounting frustration, particularly in the area of business data, that in many case, they are providing protection to information, that, by any reasonable standards, do not appear to need to be protected from disclosure because it is considered to be public information or not sensitive in nature.

Another exception to the rule of secrecy is data-sharing. In 1971, the mandate of Statistics Canada was expanded to include the promotion of avoidance of duplication in the information collected by departments of government. The mechanism to carry out this part of the mandate is the provisions in the Act to enter into data-sharing agreements with any other department, municipal or other corporation. When a data-sharing agreement is struck, the legislation requires that respondents be informed of the agreement and that they be given the right to object to the sharing of their information.

3. ACCESS TO INFORMATION ACT

The *Access to Information Act* gives the public the right of access to documents held by federal government institutions, with certain limited and specific exemptions. One of these exemptions expressly prohibits third-party access to information collected pursuant to the *Statistics Act* which could identify a respondent. Moreover, all of the information which Statistics Canada puts or can put at the disposal of the public is excluded from the *Act*.

4. PRIVACY ACT

Last, but definitely not least in the triumvirate of legislation governing Statistics Canada activities is the *Privacy Act*. It contains explicit requirements pertaining to collection of personal information. More specifically,

- government institutions may collect personal information only if it relates to their own operation programs or activities,

- individuals from whom personal information is collected must be informed of the purpose to be served by the collection and

- government institutions may use personal information only for the purpose for which it was collected or for a consistent purpose unless the individual gives consent to another use.

The *Privacy Act* also gives the authority to government institutions covered by this legislation to disclose personal information without consent of the individuals if another Act of Parliament such as the *Statistics Act* authorizes its disclosure. This allows our continuing access to administrative records containing personal information. These provisions are also found in provincial privacy legislation.

5. INTERRELATIONSHIP

From STC's point of view, the three pieces of legislation work well together. From a collection perspective, since data are gathered for statistical purposes only and since personal as well as other identifiable information is protected from access and disclosure, compliance with the *Privacy Act* has not required major adjustments.

In carrying out its dissemination role, the Agency's emphasis is in full congruence with the basic thrust of the *Access to Information* and *Privacy Acts* in that it provides maximum availability of statistical information at levels of detail which do not disclose individually identifiable data provided by the respondents.

Of the two Acts, the *Privacy Act* has had more of an impact on how we carry out our business. In fact, two of the internal policies that are part of our legislative/policy framework have been developed specifically to deal with privacy concerns.

6. RECORD LINKAGE POLICY

Departmental policy covering data linkage was developed almost ten years ago in response to concerns expressed by the public and the Office of the federal Privacy Commissioner regarding the possible linkage of personal information from a multitude of sources, actualized as a result of technological progress, without the individuals concerned being aware of it.

Although the concerns expressed were focused primarily on linked information being used for administrative or regulatory purposes, Statistics Canada, very much aware of the importance of the data linkage technique for a statistical agency, wanted to ensure that its utilization for statistical and research purposes would not be unduly restricted, or worse, prohibited.

STC record linkage policy permits this activity only if it satisfies numerous conditions, including the following:

- the linkage must be used for statistical or research purposes consistent with STC's mandate;

- dissemination of linkage products must satisfy *Statistics Act* confidentiality provisions;

- the linkage must not be used for purposes that would be detrimental to the respondents involved; obviously, the resulting benefits must serve the public interest; and

- the linkage must be consistent with a prescribed review and approval process.

The review process is multilayered. The first layer is a senior management committee which analyzes all linkage proposals to formulate a recommendation for the Policy Committee, which is made up of the Chief

Statistician and Assistant Chief Statisticians. If the Policy Committee supports the recommendation of the Junior Policy Committee, the Chief Statistician submits the proposal to the Minister responsible for Statistics Canada. Parenthetically, depending on the sensitivity of the linkage, the Chief Statistician may decide to expand the review policy to include consultation with external agencies.

7. POLICY ON INFORMING SURVEY RESPONDENTS

Another policy that addresses privacy concerns is the *Policy on Informing Survey Respondents*, the thrust of which is to apprise respondents of the reasons for taking a given survey. By doing this, Statistics Canada not only meets the requirements of the *Privacy Act* when collecting personal information, it also encourages the cooperation of respondents since informed respondents are more apt to collaborate if they understand the purpose of an information collection. They are also more likely to provide better quality information.

All survey instruments used by STC in its collection activities are reviewed centrally. In the review process statements of use and purpose of the surveys are assessed as well as statements assuring respondents of confidentiality protection. The review also ensures that all legal requirements are met, that surveys are properly authorized and that when a voluntary survey is undertaken, an enabling order is obtained.

A survey submitted for review is also scrutinized from a privacy perspective e.g. if identifiers are retained after collection, a personal information bank must be created and registered. This conforms with a requirement of the *Privacy Act* to identify personal information banks and to have those banks described in an index of personal information which is published by Treasury Board.

In occurrences where an information collection raises a number of privacy issues, a meeting will be arranged with officials of the Privacy Commissioner's Office. The purpose of the meeting is not to obtain their seal of approval since it is understood that they have to retain their impartiality but more to brief them and to obtain their views on whether STC has covered all the bases in complying with the provisions of the *Privacy Act*.

8. MICRODATA RELEASE POLICY

The implementation of the policy on release of micro-data has been assigned to the Microdata Release Committee which is probably the longest-standing STC committee concerned with confidentiality of information. It was created in the early 1970s when the *Statistics Act* was revised and one of the changes made allowed the release of anonymous individual information.

Under the current policy, releases of micro-data are authorized only if they substantially enhance the analytical value of the data collected and when the Agency is satisfied that all reasonable steps have been taken to prevent the identification of particular survey units. To minimize the risks of disclosure, the policy requires that all micro-data files considered for release be reviewed by a group of experts who use established criteria and their judgement in the assessment of files.

Screened microdata files of household data have been released for some time now. There is increasing pressure to produce this form of output, and to include more detail on them. Concurrently, users are becoming more sophisticated and access to more powerful hardware and software is making it easier to potentially link the files to other survey and administrative files. Compounding the problem are some of the new longitudinal surveys that have been undertaken by Statistics Canada. In principle, the risk of identification from longitudinal data increases the potential risk of disclosure. However, microdata are the only form of output that can properly exploit the potential of longitudinal surveys and these surveys were justified on the expectation that their analytical potential could be realized by a wide variety of users.

This is an issue that is causing much concern and is currently under discussion by members of the Microdata Release Committee and the Confidentiality and Legislation Committee. It is also the subject of active methodological research.

9. SECURITY POLICIES

The main objective of STC's established security policies is the protection of all sensitive information including machine-readable information from unauthorized access.

There is a growing concern with respect to the real capacity to protect computerized information. Media reports of hackers breaking into supposedly well protected computer systems have made the public more sceptical about assurances that information provided is secure.

The basic feature of our EDP policies is the requirement that sensitive statistical information be processed, stored and transmitted only on a network to which public access is not allowed.

Our security policies also require that sensitive statistical information must be controlled at all times. Only authorized employees with a need to know can access this information, an audit trail on the movement of that information within STC is mandatory and micro-information cannot be removed from the workplace unless specific dispensation is obtained from the Chief Statistician.

10. CONCLUSION

It would seem that the legislative/policy framework has been an effective tool. There is a very clear relation between STC's ability to ensure confidentiality of information collected, to be sensitive to the privacy rights of respondents, to provide appropriate security measures and the success the Agency has in obtaining information from the public.

This has been substantiated over the years by good response rates to our surveys. It has been further substantiated by the results of a survey on privacy undertaken three years ago. Even though 92% of survey respondents indicated at least a moderate level of concern about privacy only 14% expressed concern about providing personal information to Statistics Canada.

We know that concerns about privacy will continue to grow as the collection, accumulation and use of personal information flourishes. In order to keep on providing the information a society must have to effectively address its economic and social problems, it is essential that the climate of trust that currently exists between STC and its respondents is nurtured. Not only is this a challenge of some dimension it is also a responsibility

SESSION 4

Quality of Statistical Information

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

TAKING UNCERTAINTY AND ERROR IN CENSUSES AND SURVEYS SERIOUSLY

S.E. Fienberg[1]

ABSTRACT

Government statistical agencies release extensive amounts of micro-data from censuses and surveys but the users of such data releases are often perplexed as to how to use "sampling weights" and other information about uncertainty and error which the agencies provide. Sampling weights represent only one component of such uncertainty and error, and they play different roles in analyses depending on the perspective of the user. Nonsampling errors are typically of greater concern yet play a diminished role in agency reports and users' modelling. Finally, agency-injected error to preserve confidentiality represents one further level of uncertainty. One way to get both the agencies and the users to take all of these levels of uncertainty and error seriously is to think about the reporting of microdata in a new way. This paper outlines a new integrated approach to the release of micro-data and the reporting of uncertainty and error, consistent with modern statistical methodological practice, that should allow both agencies and users to take the error and uncertainty in census and survey data seriously.

KEY WORDS: Bootstrap; Confidentiality; Contingency tables; Cumulative distribution function; Data disclosure avoidance; Loglinear models; Multiple imputation; Regression models.

1. INTRODUCTION

1.1 Goals

In moving from raw census and survey data to information, the theme of this symposium, we often speak about the need to separate the signal from the noise. Understanding the implications of uncertainty and error is crucial to this task and survey statisticians have developed an elaborate set of tools for examining error and variability and for using this information to shape the limits of inference about various underlying social phenomena. But the job becomes much more complex when the goal of a statistical agency is to share information in the form of one or more data releases that are intended to enable a wide variety of users to analyze the released data and to move from these data through information to wisdom. A key argument advanced in this paper is that the broad use of statistical methodology, in an integrated fashion, can facilitate the attempt to achieve wisdom.

The producers of census and survey data, the statistical agencies, often lament the fact that the users of the data fail to pay attention to the information on uncertainty and error which they regularly provide. The users, in turn, argue that the agencies fail to take their analytical goals and concerns into account. Getting users to take variability seriously, the essence of the title of this paper, thus involves bridging this divide. The core of the paper argues that to get users to take uncertainty and error seriously we need to take a new integrated approach to the entire survey uncertainty and error enterprise, from collection through editing and data disclosure avoidance adjustments, via a model-based estimation approach. While adopting such an approach will not solve all problems, it might well eliminate some of the most problematic aspects of survey weighting and survey analysis. The approach we suggest would represent a radical departure from current agency practice, but would have the salutary effect of integrating modern survey approaches with mainstream statistical methodology as it is practiced outside the survey realm.

The goal of this paper is to address the implications of uncertainty and error in the context of censuses and surveys from three different perspectives: that of the producers (e.g., statistical agencies), that of the users, and that of statistical methodologists. In the process, we

[1] Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890.

provide commentary on current practice and suggest methodological strategies for the development of a unified framework to address uncertainty and error.

We begin with two assertions:

- Government statistical agencies are doing an excellent job of census/survey design, data collection, and measurement, including the documentation of the various sources of survey error.
- The users of government statistical data, both academic users and those using statistical data for policy purposes, tend to have well-defined analytical objectives, and they are more than willing to reflect the uncertainty associated with the data in addressing these objectives.

1.1 The Problem

If statistical agencies are doing such a good job and the users are so willing, what is the problem? To our mind one of the problems is that survey statisticians are often left to define the scope and form of substantive problems and what measures of them are important. As Fred Gault and Martin Wilk (1995) recently suggested in the context of science and technology statistics for Canada:

> Like it or not, statisticians and accountants specify measures and indicators that often preempt the arenas of policy focus. . . [that is] measurement schemes implemented by statisticians may have a profound influence on [policy, and this work] ... is far too important to be left to the statisticians.

Traditionally the survey statisticians have thought in terms of general goals for survey data collection and descriptive uses of survey data. They have developed statistical methodologies that are in accord with this perspective. The users, on the other hand, more often than not think in terms of focused analytical and policy goals, and they typically have in mind a framework that fits with standard statistical models of a regression-like variety. Such models have there own inherent variability and error built into them, but do not necessarily reflect the sample survey features that were important in the data collection. As a consequence, there is a mismatch, or at least a gap, between what agencies provide in way of information on uncertainty and error for the data that they release, and what the users would like to do with the released data, including their capacity to cope with the information on uncertainty and error.

To resolve this apparent mismatch, we suggest that both parties need do things differently! Agencies need to report data and information in a different form and users need to have a systematic approach to fully reflect the inherent uncertainty and errors in their analyses of released data. Our prescription builds on recent developments in statistical methodology in order to integrate sources of error at agency level, and to provide users with data in form that makes it easy to take error into account in modeling and inferences. The framework for our prescription is still in its formative stages, and many technical details require attention. Nonetheless, we believe that a discussion of these issues is timely and that the current version of the framework does provide the basis for first steps towards a unified statistical approach to data collection, data release, and data analysis.

1.2 Organization of Paper

The remainder of the paper is structure as follows. In Section 2, we outline in a somewhat more detailed form the contrary positions of the producers and the users. Then, in Section 3, we present the elements of our unified framework including a new approach to data disclosure avoidance and the release of public use micro-data files. In section 4, we explain how this new approach allows the users to take error and uncertainty more seriously and we point out several key research questions, the answers to which are crucial for real-world implementation. In an Appendix, we outline the relationship between some popular disclosure avoidance procedures for sets of categorical variables and the class of loglinear models, and we sketch some of the elements of a model-based implementation of the strategy proposed in this paper.

2. PERSPECTIVES ON SOURCES OF ERROR AND THE ANALYSIS OF SURVEY DATA

2.1 The Agency's Perspective

What I use to characterize the agency perspective begins with the traditional approach to survey design and measurement as described in books and articles on sampling and nonsampling error, and then proceeds through the key elements of data processing and release that are part of agency practice. The sources of error from this perspective typically include the following elements:

- Frame error (e.g., differential census undercount);

- Sampling error (complex survey design);
- Nonresponse error (bias and variability);
- Editing error (e.g., imputing missing values);
- Matching error (for merged data files);
- Other nonsampling error (e.g., mode of interview, questionnaire design, etc.);
- Confidentiality edit error (e.g., due to error injected into the data as a result of top-coding, cell suppression, added noise, data swapping).

The standard approach to such an array of errors and uncertainty is one of "divide and conquer," with agencies addressing each component or even subcomponent, but almost always separately from the rest. We do not use the phrase "divide and conquer" here in any pejorative sense, but rather as a reflection of the need that government statistical agencies have to get on with the tasks at hand, yet take seriously the multiplicity of problems beset real survey data (c.f. Patz 1996, Fienberg, Gaynor, and Junker, 1996). Once each component is conquered, statisticians need to focus on how to put it together with other components in an integrated form (e.g., see Groves, 1989; Lessler and Kalsbeek, 1992). It is rare to find an integrated model for error that can actually be used for analytical purposes.

2.2 The Users' Perspective

Typical users of government statistical data are interested in relationships and causal connections for policy choices. They use statistical models to describe such relationships. Often their view of "error" is akin to including an error component in an analytical model (e.g., such as a regression error term ϵ in the equation $Y = b_0 + b_1X+\epsilon$). Otherwise, the typical user has limited ways to address the multiplicity of information on uncertainty and error coming from the statistical agency that produces the data.

For decades, samplers and survey statisticians (with some notable exceptions) have attempted to convince social scientists and government policy makers that the objectives for most government sample surveys and censuses were descriptive rather than analytic (for an early discussion of the differences see Deming, 1978). One of the consequences of the perpetration of this "descriptive" myth has been a gulf between what the survey takers and government agencies produce, on the one hand, and what the users of their products attempt to do with released data, on the other hand. Survey sampling weights (usually reflecting the probabilities of selection, as well as selected non-response adjustments) and instructions on sample variances typically provide the interface between the producers and users, with some naive users doing weighted analyses simply because the weights are reported on the released tape and because they believe (erroneously) that this is the correct way to take into account the eccentricities of the survey design in their model-based analyses.

2.3 Articulating the Users' Objective

As we noted above, the typical user is interested in analytical models and especially ones with causal implications. Thus we can think of the users' objectives as involving the linking of response variables, Y, and explanatory variables, X, through a statistical model that attempts to represent some underlying substantive phenomenon. Unfortunately we rarely get to observe or measure Y and X directly. What is produced through a census or a survey questionnaire is often a related but fallible measure of the quantities of real interest. These we label Y^* and X^*.

The user is interested in models for the conditional distribution of Y given X and thus we can take as the user's objective the estimation of a multivariate cumulative distribution function (c.d.f.), of the forms $F_{X|Y}$ or $F_{Y|X,\theta}$ for various values of X, or at least characteristics of such a multivariate c.d.f. Here the parameter θ might be a population mean or variance, μ or σ^2 or a parameter(s) in a statistical model such as a regression coefficient, β, likely multidimensional in form. While there has been some interest in the survey literature in the problem of estimation distribution functions (e.g., see Rao, 1994, and the references contained therein), although this literature has been concerned primarily with univariate Y. In the ensuing discussion we ignore those sources of measurement error in X beyond those forms captured in the agency's own evaluation and data preparation activities.

Estimation of a multivariate c.d.f. is a general statistical problem that includes a number of interesting special cases. For example, suppose that all of the variables in the user's model and in the data set are categorical in nature, as is often the case in censal and survey settings. Then the c.d.f. is essentially equivalent to the table of conditional probabilities (for Y given X) that correspond to the cross-classification of the variables in contingency table form (c.f., Bishop, Fienberg, and Holland, 1975). We refer to this special case again in the Appendix and provide an extended set of references and notes on this special case. Fienberg , Makov, and Steele (1996) provide further details.

2.4 The Current Agency Approach

At the risk of oversimplification, we can characterize the standard approach to data collection, processing and release roughly as follows:

- Collect and "clean up" the raw data. This includes editing, matching and all other preliminary processing.

- Protect the data by applying some form of data disclosure avoidance methodology.

- Then release the resulting data in one or perhaps both of the following forms:
  – as set of marginal tables for some larger cross-classification (i.e., selected marginal cross-classifications - - see the discussion in the Appendix regarding the relationship between marginal tables and loglinear models).
  – as micro-data files for the variables related to the ones of user interest (Y^*, X^*).

- Estimate θ directly using a sample-based quantity, $\bar{\theta}$.

In effect, the user then follows the agency's lead and estimates the c.d.f., $F_{Y^*|X^*}$ or $F_{Y^*|X^*,\theta,}$ directly from the released data using the "empirical" c.d.f. (suitably weighted to take into account the impact of the survey design), $\bar{F}_{Y^*|X^*}$, or possibly a more elaborate and smoother parametric estimate based on the estimated parameter, $\bar{\theta}$ i.e., $\bar{F}_{Y^*|X^*,\theta}$.

2.5 Shortcomings of The Current Approach

While this approach might make considerable sense for some descriptive statistical problems, the fact is that $\bar{F}_{Y^*|X^*}$ and $\bar{F}_{Y^*|X^*,\bar{\theta}}$ rarely reflect fully aspects of sampling design error that many believe to be important, such as clustering, and they almost never reflect the other sources of error listed above that typically dwarf sampling error. Further, given the relatively primitive statistical state of disclosure avoidance methodology, the user may still be able to "identify" individuals in the released data. One way to overcome these shortcomings is to continue to address the various components of error and to separately improve the approach to data disclosure avoidance. Alternatively, we can attempt to reconceptualize the data reporting problem in a new and integrated fashion.

3. A NEW STRATEGY AND FRAMEWORK

In this section, we propose a new approach to the release of survey data. We begin with the goals of the users and ask how agencies should organize the data of interest in order to provide data releases that fit with the users goals.

3.1 Generating "Pseudo" Micro-Data Files for Public Use

Our new approach is cast in terms of the release of a public-use micro-data file that is intended to support analyses for the conditional distribution of Y given X. the first step in our prescription is:

1. Combine the census or survey data that the agency would normally have chosen to release, in form $\bar{F}_{Y^*|X^*}$ and $\bar{F}_{Y^*|X^*,\bar{\theta}}$, with formal statistical information on error, e.g., from editing, matching, nonresponse, etc., and apply some form of parametric or semi-parametric technique to estimate $F_{Y|X}$ and $F_{Y|X,\theta}$ by $\hat{F}_{Y|X}$ and $\hat{F}_{Y|X,\theta}$ respectively, where θ is a new estimate of θ cast in terms of the distribution of the variables of actual user interest, Y and X.

For non-parametric estimation of F, we can either think in terms of a classical statistical approach using some type of kernel density estimator or a related type of "smooth" estimate (e.g., see Scott, 1992), or a Bayesian approach based on the mixture of Dirichlet processes (e.g., see West, Müller, and Escobar, 1994; Gelfand and Mukhopadhyay, 1995) or the use of Polya trees (Lavine, 1992). These tools, however, have been used primarily in low-dimensional problems and thus there needs to be additional research to study their adaptation to the high-dimensional censal and survey problems which are the focus of this paper. Even if these methods are not especially efficient for statistical estimation purposes, they may serve the needs of data disclosure avoidance which are crucial to the strategy outlined here.

In what ways this new smoothed estimate of $F_{Y|X}$ differs from the one that is explicit or implicit in the current approach? We offer three examples. First, consider the release of census data. In both the US and Canada, there has been extensive documentation of the extent of census undercoverage and how the resulting undercount is distributed across groups in the population and across geographic areas. Failure to correct for such undercoverage in the release of data of the form $\bar{F}_{Y^*|X^*}$ leads to biased estimates of the true quantity of interest, $F_{Y|X}$. Second, by smoothing data to reflect regression-like relationships we can typically achieve improved estimates with much lower variances, although at the price of some potential bias. Finally, by incorporating agency information on components of error (which tends to increase variances) into the statistical estimation process, we produce a new smoothed estimator of $F_{Y|X}$.

The next steps in our prescription are:

2. Instead of releasing the c.d.f. estimated in step 1 above, the agency now "samples" from it to create a "pseudo" micro-data file which we label as $\hat{\bar{F}}_{Y|X}$ and $\hat{\bar{F}}_{Y|X,\theta}$. (We use the overbar to indicate a sample from the smoothed c.d.f.'s, in accord with our earlier notation for the empirical c.d.f, which corresponds to a sample).

3. The agency repeats the process of "sampling" and then releases the resulting replicate "pseudo" micro-data files.

3.2 Features of Pseudo Micro-Data File

The "pseudo" micro-data files created in the approach outlined above have several interesting features. First, if we think of $\hat{\bar{F}}_{Y|X}$ and $\hat{\bar{F}}_{Y|X,\theta}$ as consisting of a set of released records for individuals, then the these "individuals" do not necessarily correspond to any of those individuals in original sample survey. This enhances the public notion of the protection of confidentiality of responses even if an intruder might still be able to indirectly make inferences about individuals in the original sample.

This point is especially important from the perspective of data disclosure avoidance. Since the individuals in the pseudo micro-data file are not necessarily those from the original sample, we have at least in part addressed confidentiality concerns. After all, we no longer even appear to be releasing data for any individual from the original sample. But this discussion of data disclosure avoidance is somewhat illusory. It remains possible that individuals, whose values on Y and X are far from those for the rest of the sample, may still in effect be regenerated through this complex statistical estimation process and reemerge virtually intact in the pseudo micro-data file. Thus we would argue that empirical checks on the effectiveness of data disclosure avoidance are still necessary and, in particular, we would advocate examining the issue from the perspective of an intruder (e.g., see Fienberg, Makov, and Sanil, 1994).

Second, there is close connection here with two recently developed statistical methods: (1) the bootstrap (Efron 1979, Efron and Tibshirani 1993, Hall 1992) which is a classical method involving repeated sampling (with replacement) from an empirical distribution function; (2) multiple imputation (Rubin 1987, 1993) which is a Bayesian method for generating values that are sampled from a posterior distribution. Our preference is to think about the estimation implicit in the approach outlined here from a Bayesian point of view. Thus, in effect, we are proposing that agencies should first estimate the empirical distribution function, generating the full posterior distribution of $F_{Y|X}$ or $F_{Y|X,\theta}$ and then sample from it using Rubin's multiple imputation approach. From this perspective, the bootstrap can be viewed as a way to sample from something approximately akin to the mean of the posterior distribution.

Third, the sample design for the released records need not be same as that for original sample survey. Thus, at least in principle, the agency could use simple random sample or even sampling with replacement from $\hat{F}_{Y|X}$ or $\hat{F}_{Y|X,\theta}$ Rubin (1993) emphasizes this point without explaining exactly how to determine what we might call the "equivalent" sample size for the released data files. The heuristic idea is that there is only so much in formation available in the data and the resampling process cannot increase this. To preserve the appropriate level of accuracy in the data we need to have a bootstrap sample size that at least is conceptually equivalent to the "effective sample size" of the complex sample design, thus reflecting a design effect. This notion is somewhat problematic, however, as the "effective sample size" might well vary from one analytical setting to another!

But perhaps the most important feature of the approach is that users can now analyze pseudo micro-data files to estimate specific quantities of interest, e.g., θ, using *standard statistical methodology*. In essence the idea is that we can use a standard statistical method such as regression analysis or something more elaborate and thus will produce consistent estimates of the coefficients of interest. What we cannot do, however, is use the usual estimates of standard errors that result from the standard analysis tools. One of the lessons from both the bootstrap and multiple imputation is that while we can estimate θ using standard statistical methodology applied to the generated bootstrap or multiple imputation sample, we cannot get a proper handle on the variability of our estimates without using replicate versions of the pseudo micro-data file. Generating multiple replicates, however, is a relatively simple task and estimating variances using the multiple versions of estimated parameters is then straightforward and requires no special computer programs.

4. TAKING VARIABILITY SERIOUSLY

We believe that it is important for us to distinguish between the idea of generating public-use micro-data files based on real people and real data through a statistical simulation process, such as we have outlined in this paper, and the typical micro-simulation model, which may rely on related statistical models but which does not correspond to data on real people. There is a serious difference between "pseudo people" who resemble individuals from whom we have actually collected data of interest, and "imaginary people for whom we have invented data through a stochastic or nonstochastic modelling process. In this paper we propose the former, not the latter.

4.1 Virtues of Proposed Framework

There are several virtues of the proposed framework outlined above. First, we believe that it would force agencies to take their own data and their sources of error more seriously, as these are key inputs to the modeling effort outlined in Section 3. Second, we believe that it would solve a large part of the data disclosure avoidance problem. Third, the framework would generate public-use micro-data files of a form that would allow users to apply standard statistical methodology and model search methods. All of these benefits or virtues would thus move both parties toward more effective and simple-to-use method of variance estimation, thus addressing the title of this paper.

4.2 Examples of Research to Be Done

There are a number of formidable technical details that need to be addressed before an agency could properly implement the proposed framework. Examples of these include:

- How should an agency combine the multiple sources of error and uncertainty?

- What smoothing methods should be used and how much smoothing is appropriate?

- How do we determine "effective" sample size for pseudo micro-data files? The application of bootstrap ideas relies on certain series expansions (e.g., see Hall, 1992), and these typically require the use a bootstrap sample of the same size as the original sample. What is the equivalent notion here?

- How many replicates are required for variance estimation? Rubin (1987, 1993) suggests the use of four or five replicates in the multiple imputation context. Efron and Tibshirani (1993) uses large numbers of bootstrap replications. Will a smaller number suffice for either approach?

Further the actual implementation of algorithms for the highly multidimensional situations involved in censal and survey data may require new statistical methods and theory. For example, as we suggest in the Appendix, the problem of simulating from distributions for multidimensional contingency tables subject to marginal constraints has been implemented primarily for two and three-dimensional tables. Implementation for higher dimensions requires new strategies and algorithms. These are at the forefront of current statistical and mathematical research.

Finally, we may need to think about the statistical estimation problems outlined hear in a form different from that which we usually find in the methodological literature. Be cause of the multiplicity of goals that we are attempting to address, we may need to think in terms of providing the users with data that enable them to approximate the conditional distributions $\hat{F}_{Y|X}$ and $\hat{F}_{Y|X,\theta}$ rather than reproduce them in a more precise statistical fashion.

4.3 Summary

In this paper, we have tried to suggest that both government agencies and users bear responsibility when it comes to utilizing census and survey data. It is no longer enough for agencies to prepare public-use files and extensive sets of tabulations as they have in the past. Nor can they continue to ignore the analytical goals of the users of their data. At the same time, the users must learn how various sources of survey error affect their analytical goals, and to build such information into the statistical procedures they use.

We have argued that, by looking to and utilizing recent developments in statistical methodology, we may be able to develop an integrated approach to the release and analysis of survey data which will help us all learn to take uncertainty and error seriously. Perhaps the framework proposed in this paper will be the first step towards this goal.

APPENDIX: NOTES AND REFERENCES FOR THE CATEGORICAL DATA CASE

This appendix provides an annotated outline of the estimation and simulation process of Section 3 for the special case of categorical variables and cross-classifications. Our focus is on parametric estimation of the c.d.f. which as we note above in equivalent to estimating the cell probabilities in a contingency table. Fienberg, Makov, and Steele (1996) provide further details and description.

The most common class of statistical models used in connection with contingency table data is the loglinear model and for a set of basic sampling schemes (e.g., see Bishop, Fienberg, and Holland, 1975) there is a direct relationship between a specific hierarchical loglinear model and a set of marginal tables that correspond to the minimal sufficient statistics associated with the model. If we report only those marginal totals appropriate for a log-linear model that fits the data well, then another investigator can, in effect, reconstruct the cell probabilities for the full contingency table (c.f., Fienberg, 1975). Further, reporting only a specific set of marginal tables is saying that these are the only totals needed for inference and thus is implicitly suggesting the appropriateness of a specific log-linear model.

The two most commonly used methods for data disclosure avoidance in categorical variable settings are (i) cell suppression (e.g., see Carvalho, Dellaert, and Osório, 1994; Cox 1980, 1995; Robertson, 1993; and Subcommittee on Disclosure-Avoidance Techniques, 1994) and (ii) data swapping (e.g., see Dalenius and Reiss 1982; Griffin, Navarro, and Flores-Baez, 1989; and Subcommittee on Disclosure-Avoidance Techniques, 1994). Unfortunately there seems to be a total disconnect between the literature on disclosure avoidance for categorical variables and the now standard literature on loglinear models for categorical data. This is rather unfortunate since the notion of margin preservation is fundamental to both cell suppression and data swapping. In the former, cells are suppressed subject to marginal constraints and, in the latter, individuals with one set of margins fixed are swapped between cells thus preserving other totals. Thus key features of these methods can be embedded in the loglinear model framework thus suggesting alternative ways to approach disclosure avoidance. Further results from the log-linear model literature may well be of value in understanding the properties of methods such as cell suppression and data swapping (c.f. the discussion in Fienberg, 1995).

Finding a cross-classified table of counts that satisfies a given set of marginal constraints is a problem which has occupied the attention of a substantial number of statisticians in recent years (e.g., see Agresti, 1993; Zelterman, Chan, and Mielke, 1995). An number of algorithms have been proposed but they have been implemented primarily for two- and three-way cross-classifications. New ideas from the literature on graphical loglinear models suggest that implementation for higher dimensions may at last become feasible (e.g., see Diaconis and Sturmfels, 1993 for a proposed algorithm and Lauritzen, 1996 or Whittaker, 1990 for details on graphical models). The framework we outline in Section 3 requires us to produce a smooth c.d.f., and then sample from it. In the present context, this seems to suggest, at least heuristically, that we should consider making draws from the exact distribution conditional on a fixed set of marginal totals. But that we might also choose to use only those a generated table if it at least satisfies some higher-order loglinear model importance of graphical models Lauritzen and Whittaker, Diaconis and Sturmfels. Alternatively we can generate a full posterior distribution of the cell probabilities in the table, e.g., using the methods of Epstein and Fienberg (1992), and then sample from that posterior distribution.

For further development of the issues and approaches described in this appendix, see Fienberg, Makov, and Steele (1996).

ACKNOWLEDGMENTS

The preparation of this paper was supported in part under a contract with WESTAT and the U. S. Bureau of the Census. I am especially appreciative to David Binder for initial reactions to the general strategy to Peter Müller, Danny Pfefferman, and Steffen Lauritzen for providing references that have found there way into this paper. None of them, however, bear any responsibility for my use of the suggested materials, or for my admittedly speculative ideas on the applicability of specific statistical methods.

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion), *Statistical Science*, 7, 131-177.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice, Cambridge, MA: MIT Press.

Carvalho, F.de, Dellaert, N., and Osório, M.deS. (1994). Statistical disclosure in two-dimensional tables: General tables, *Journal of the American Statistical Association*, 89, 1547-1557.

Cox, L. (1980). Suppression methodology and statistical disclosure control, *Journal of the American Statistical Association*, 75, 377-385.

Cox, L. (1995). Network models for complementary cell suppression, *Journal of the American Statistical Association*, 90, 1453-1462.

Dalenius, T., and Reiss, S.P. (1982). Data-swapping: a technique for disclosure control, *Journal of Statistical Planning and Inference*, 6, 73-85.

Deming, W.E. (1978). Sample surveys: the field, in *International Encyclopedia of Statistics, Vol 2.* (W.H. Kruskal and J.M. Tanur, eds.), New York: Macmillan and the Free Press, 867-885.

Diaconis, P., and Sturmfels, B. (1993). Algebraic algorithms for sampling from conditional distributions, Unpublished manuscript.

Efron, B. (1979) Bootstrap methods: another look at the jackknife, *Annals of Statistics*, 7, 1-26.

Efron, B., and Tibshirani , R. (1993). *An Introduction to the Bootstrap,* New York: Chapman and Hall.

Epstein, A.D., and Fienberg, S.E. (1992). Bayesian estimation in multidimensional contingency tables, *Proceedings of Indo-U.S. Workshop on Bayesian Analysis in Statistics and Econometrics* (P.K. Goel and N.S. Iyengar, eds.), Lecture Notes in Statistics Vol. 75, New York: Springer-Verlag, 27-47.

Fienberg, S.E. (1975). Perspectives Canada as a social report, *Social Indicators Research*, 2, 153-174.

Fienberg, S.E. (1994a). Conflicts between the needs for access to statistical information and demands for confidentiality, *Journal of Official Statistics*, 10, 115-132.

Fienberg, S.E. (1994b). A radical proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality, Technical Report No. 611, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Fienberg, S.E. (1995). Discussion of presentations on statistical disclosure methodology, *Seminar on New Directions in Statistical Methodology, Statistical Policy Working Paper No. 23,* Federal Committee on Statistical Methodology, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, Part 1, 68-79.

Fienberg, S.E., Gaynor, M., and Junker, B. (1996). Discussion of "Modelling mortality rates for elderly heart attack patients: Profiling hospitals in the Cooperative Cardiovascular Project," in *Case Studies in Bayesian Statistics III*, New York: Springer-Verlag (in press).

Fienberg, S.E., Makov, U.E, and Sanil, A. (1994). A Bayesian Approach to data Disclosure: Optimal Intruder Behavior for Continuous Data. Technical Report No. 608, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Fienberg, S.E., Makov, U.E., and Steele, R. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and loglinear models. *Proceedings of the Annual Research Conference, U. S. Bureau of the Census*, U. S. Department of Commerce, Washington, DC, (to appear).

Gault, F.D., and Wilk, M.B. (1995). Science and technology measurement -- Rhetoric and reality, paper prepared for presentation at "International Symposium on Measuring R&D Impact," Ottawa, Canada, September 13-15, 1995.

Gelfand, A.E., and Mukhopadhyay, S. (1995). On nonparametric Bayesian inference for the distribution of a random sample, *Canadian Journal of Statistics*, 23, 411-420.

Griffin, R., Navarro, A., and Flores-Baez, L. (1989). Disclosure avoidance for the 1990 census, *Proceedings of the Section on Survey Research,* American Statistical Association, 516-521.

Groves, R.M. (1989). *Survey Errors and Survey Costs*, New York: John Wiley.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Lessler, J.L., and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys,* New York: John Wiley.

Lauritzen, S. (1996). *Graphical Association Models,* New York: Oxford University Press.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling, *Annals of Statistics*, 20, 1222-1235.

Patz, R. (1996). Hierarchical models for new modes of educational assessment, unpublished Dissertation, Department of Statistics, Carnegie Mellon University.

Robertson, D. (1993). Cell suppression at Statistics Canada, *Proceedings of the Annual Research Conference, U. S. Bureau of the Census*, U. S. Department of Commerce, Washington, DC, 107-131.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D.B. (1993). Discussion, statistical disclosure limitation, *Journal of Official Statistics*, 9, 461-468.

Rao, J.N.K. (1994). Estimation totals and distribution functions using auxiliary information at the estimation stage, *Journal of Official Statistics*, 10, 153-165.

Scott, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization,* New York: John Wiley.

Subcommittee on Disclosure-Avoidance Techniques (1994). *Report on Statistical Disclosure Methodology*, Statistical Policy Working Paper No. 22, Federal Committee on Statistical Methodology, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC.

West, M., Müller, P., and Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation, in *Aspects of Uncertainty* (P.R. Freeman and A.F.M. Smith, eds.), New York: John Wiley, 363-386.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics, New York: John Wiley.

Zelterman, D., Chan, I.S.-F., and Mielke, P.W., Jr. (1995). Exact tests of significance in higher dimensional tables, *American Statistician*, 49, 357-361.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

PROBLEMS OF RESOURCE ALLOCATION

T.M.F. Smith[1]

ABSTRACT

Resource allocation is a problem which has received little attention in the literature. The discrete decision theory framework of Felligi and Sunter (1973), as implemented by Linacre and Trewin (1993), offers a way forward. Various loss functions are discussed and the total *MSE* is proposed as a suitable measure of total quality. The procedure can only be implemented if estimates of variances, biases and of marginal costs can be made in numerical terms, using, if necessary, professional judgement. The formal framework focusses attention on the areas where more information is needed.

KEY WORDS: Decision theory; Loss functions; Survey errors; Survey costs.

1. INTRODUCTION

Oskar Morgenstern (1963), in his book on the Accuracy of Economic Observations, which should be compulsory reading for all who produce and use economic statistics, demonstrates how spurious are most of the measures of accuracy attached to official statistics. The quality of official statistics has been of concern since the earliest days of sampling, and the ISI report on the representative method, ISI (1926), contains the following paragraph which could serve as a mission statement for any statistical agency:

> *"The greatest importance must be attached to the existence of a state of mutual confidence between the institution which exercises the official statistical service and the population which both supplies the material for the statistics and for whose sake all the work is done. The official statistics ought of course to be exceedingly cautious of its reputation, 'it is not sufficient that Caesar's wife is virtuous, all the world must be convinced of her virtue'".*

In sample surveys the ultimate virtue is accuracy, and this can be achieved only by following procedures that minimise errors for a given cost. Error minimisation in its turn has error measurement as a prerequisite, and cost control requires adequate information on the costs of the underlying procedures. With this information it is then possible to consider the efficient allocation of resources.

The ISI report includes the following caveat which will be instantly recognized by all official statisticians today:

> *"It would be imprudent to shut one's eyes to the fact that official statistics just now are faced by a period of difficulty. While constantly increasing demands are made of the statistical service, both in regard to the extension of the statistics to new domains and to the investigations going more deeply into things, public finances are everywhere so heavily strained that there will be a tendency rather to reduce than to increase the budgets of the Statistical Offices. Under these circumstances there is nothing left but to exploit every means of doing more without increasing the available personnel and economical resources."*

Plus ça change, plus c'est la même chose! Seventy years later statisticians still face the problem of how to allocate their limited resources to achieve high quality results in the presence of a variety of errors. Although the ISI report also recognized the potential importance of non-sampling errors on the accuracy of survey estimates, it was not until the 1940's that systematic studies were made in India and the USA on the effects on estimates of

[1] T.M.F. Smith, Department of Mathematics, University of Southampton, Southampton, United Kingdom, S09 5NH.

nonresponse and of response and coding errors. The development by Hansen *et al.*, (1961) of the census model for total survey error provided a theoretical framework for investigating overall accuracy. Subsequent work has refined aspects of the model, has measured individual sources of error, and has provided methods for the adjustment of analyses in the presence of specific non-sampling errors. However, there has been little research on overall accuracy and on the related problem of resource allocation.

One agency which has worked on the problem, needless to say, is Statistics Canada, and in a paper read at the ISI Session in Vienna, Fellegi and Sunter (1973) provide a framework within which resource allocation can be discussed. The meeting was chaired by Tore Dalenius who, since the 1950s, has consistently championed the study of total survey design in the context of total survey error. The other papers at the meeting, by Jabine and Tepping (1973) and Nathan (1973), also address aspects of the problem of resource allocation but do not provide a general theoretical framework. The policy statement, Statistics Canada (1987), lays down criteria for the presentation of information about errors, and Groves (1989), explores the under-researched area of survey costs. I have found only one paper, Linacre and Trewin (1993), which has addressed the resource allocation problem in practice.

2. TOTAL SURVEY ERROR

The components of total survey error are now well understood. Every part of the survey process, from concept formulation, through frame construction, to data collection and analysis, has the potential for error. Groves (1989, p.17) gives a diagram which illustrates the components of total error broken down into biases and variances. Variances can frequently be estimated from within the survey, or from special studies carried out within the survey process, but biases often require information external to the survey for their estimation. It is not surprising, therefore, that the great bulk of research has concentrated on the more easily measurable variance components. It is mainly for censuses, where there are no sampling errors, that one finds extensive studies of biases. This allocation of research effort by survey methodologists does not reflect the perceived wisdom that the contribution to total survey error of biases may be of the same order of magnitude as that of variances.

If T estimates some population value, θ say, and k sources of error have been identified, then a model for the total error of estimation can be written as:

$$T = \theta + \sum_{j=1}^{k} A_j, \tag{1}$$

where A_j is the error from source ***j***. The structure of the errors is complex, with sampling errors being conditional on the frame, non-response on the sample, response errors on the respondents, and coding and editing on the responses. Training and management may affect errors at all levels of the collection hierarchy. Despite, or possibly because of, the complexity of the error structure, the usual assumptions about the errors are:

$$\begin{gathered} E(A_j) = B_j, \\ V(A_j) = \sigma_j^2, \\ cov(A_j, A_i) = 0, \;\; j \neq i, \end{gathered} \tag{2}$$

the latter assumption being the most dubious. For any particular error we may have $\sigma_j^2 = 0$, or $B_j = 0$, but not both. For example frame errors give biases, sampling errors usually result only in variances, any biases usually being of a small order which can be ignored in this context, while response and editing errors may lead to both biases and variances. The magnitude of the errors, and their relative sizes, will depend on the particular survey. Studies have shown that in many surveys response variances can be as large, or larger, than sampling variances, and that coverage and non-response biases have the potential to swamp the variance terms. Although variances usually reduce with increasing sample size, most biases remain constant, so that the relative importance of biases tends to increase with increasing sample size. Despite this most surveys still only measure the sampling errors and rely on qualitative statements about the non-sampling variances and biases. Often reports include confidence intervals based on estimates of the sampling errors alone, which in the presence of biases can be meaningless. Honest reporting requires that an attempt be made to measure the total survey error for every survey.

The impact of errors depends not only on the particular survey but also on what is being estimated. Although the bias in an aggregate may be small, the bias in the estimate of a domain of study for the same variable may be large. Luckily the converse is not true. If the biases at the unit level are all small then the relative biases of aggregates will also be small. Jabine and Tepping (1973) point out that although the bias on a net change may be small if the survey conditions on the two occasions are similar, the corresponding biases on gross changes may be large. Biases can vary over both time and space, and they do not necessarily cancel when differences are taken, as in measures of change. The appalling failure of the public opinion polls in the UK 1992 general election has been attributed to many factors, but none of these explain the success of the polls in previous elections. The most likely explanation is that all the polls have been subject to a wide variety of errors and that in 1992 all the errors stacked up in the same direction. Opinion polls can be validated against election results. How many official surveys can be validated as rigorously? Can official statisticians be sure that the errors in their surveys don't sometimes all stack up in the same direction?

My thesis is that statisticians have failed to set up systems for the routine measurement of the major sources of survey error. Without those measurements it is impossible systematically to improve survey processes. In addition without those measurements, and the corresponding costs of survey operations, it is impossible to allocate survey resources effectively.

3. RESOURCE ALLOCATION

Resource allocation is a decision problem. Statistical decision theory starts with a loss function which measures the consequences of alternative decisions. The difficulty of defining a loss function for a complex multivariate multipurpose survey has deterred many from employing formal methods, but in any existing survey the current allocation of resources implies some judgement about the relative importance of various survey operations on survey quality. To make no change is still a decision. It seems from some of the other papers at the symposium that the allocation of resources in Statistics Canada is far from the optimum, in particular excessive resources appear to be devoted to editing in some surveys.

3.1 Loss functions

The choice of loss function should depend on the user. There are many possible users of survey data, and it would be impossible to satisfy them all. In this situation it is reasonable for a professional statistician to suggest a statistical loss function. In sample surveys, where second moments dominate, the mean square error *(MSE)* is usually chosen. The *MSE* of *T*, equation (1), is

$$\begin{aligned} MSE(T) &= E\,(T-\theta)^2 \\ &= V(T) + B^2, \end{aligned} \qquad (3)$$

where,

$$B = \sum_{i=1}^{k} B_i \ , \qquad (4)$$

is the overall bias. The *MSE* varies for every variable, and for every domain of study, and the choice of which *MSE*, or combination of *MSEs*, to use as a loss function will depend on the objectives of the users of the survey. One possible approach is to specify a range of loss functions and to evaluate the consequences of different allocations (decisions) over the whole range. The distribution of losses will then inform the final decision.

Is the *MSE* an appropriate loss function for problems of resource allocation? Although most authors use it without question there are some variations. Nathan (1973) represents the total *MSE* as the sum of a sampling *MSE* and a non-sampling *MSE*, while Fellegi and Sunter (1973) refer to their loss function as a *MSE* but in fact employ an expression of the form

$$L(T) = \sum_{1}^{k} \sigma_j^2 + \sum_{1}^{k} B_j^2 \ . \qquad (5)$$

This could be described as the total *MSE* ,rather than the *MSE* of a total, since it is the sum of the component *MSEs*.

For resource allocation decisions *L(T)* would appear to have many advantages over *MSE(T)*. Consider a survey with two errors, A_1, A_2, with variances σ_1^2 , σ_2^2 ,and biases B_1 and B_2 . If $B_1 > 0$ and $B_2 < 0$, with $B = B_1 + B_2 > 0$, then B could be reduced by increasing the size of the negative bias B_2 . In this case *MSE(T)* will be reduced whereas *L(T)* will be increased. It is difficult to argue that a procedure which has increased the bias of one error, leaving all other errors the same, has increased the total quality of the survey. This argument can be extended by recognising that errors at high levels of aggregation are sums of errors at lower levels. In the limit total quality can only be guaranteed by minimising the *MSEs* of components of error at the level of the individual unit. This can only be achieved by moving the problem upstream in the survey process and improving the procedures for data collection, measurement and processing.

Official statisticians encourage the users of their statistics to compute confidence intervals in order to take into account survey errors. The US Bureau of the Census source and accuracy statements, see Alexander (1994), are a good example. This suggests that confidence intervals could be used as loss functions. The effect of biases on the coverage properties of normal theory confidence intervals are well known. Kish (1965) has a plot of the separate tail areas and the total tail area for various nominal levels of confidence, expressed as functions of the bias ratio $R = B/\sigma$. For $R = 1.0$, and 95% confidence, the left-hand tail is 0.0015, the right-hand tail is 0.1685, with a total coverage of 0.1700. As would be expected the impact of a bias is most pronounced in the individual tails. The real problem with using the coverage level as a loss function is that the coverage can be improved by reducing the size of the ratio B/σ. This can be achieved by reducing the total bias, as above, or by increasing the variance relative to the bias. Neither represents an improvement in total quality. Coverage alone is not an adequate description of the properties of confidence intervals and is not suitable as a measure of total quality.

My conclusion is that a measure of total survey quality should be based on a sum of component measures, such as *L(T)* in (5), rather than an overall measure, such as *MSE(T)* in (3). The measure *L(T)* implicitly assumes that the component errors are uncorrelated, and Groves (1989) gives examples where this is not true. The measure could be modified to include correlations by using some form of Mahalonobis distance. An alternative to *L(T)*, which uses the original scale of measurement, is

$$L^*(T) = \sum_1^k \sigma_j + \sum_1^k |B_j| \,. \qquad (6)$$

For either *L(T)* or *L*(T)*, a reduction in any component of error leads to a reduction in loss and hence to an improvement in quality.

3.2 Feasible Actions

Once a loss function is chosen the next step in the decision process is to consider the set of feasible actions from which the final action, the decision, will be selected. An action is a change in the allocation of resources over the set of survey activities. Moving to a dual frame and simultaneously reducing sample size is a possible action. Operational statisticians, the managers of the survey process, can list the activities which can be changed, and the manner in which they can be changed. Let D_j, $j = 1$, ..., M be M identified survey activities, and let ΔD_j be the possible changes, which will usually be discrete. Activities consume resources and this consumption can be costed. Groves (1989) details cost functions for various survey activities. Let C_j be the cost of activity D_j at its present level, and ΔC_j be the change in cost of the change ΔD_j. If the level of an activity increases then the cost increases, and vice versa.

A feasible action is defined to be any set of actions that satisfies a budget constraint. For a given budget any set of changes

$$\Delta D = (\Delta D_1, ..., \Delta D_M) \qquad (7)$$

such that

$$\sum_1^M \Delta C_j \leq 0 , \qquad (8)$$

is a feasible action. So a feasible action cannot increase cost, although it could reduce cost. Operational statisticians should have information on the total costs of survey activities. For decisions they also need the marginal costs. If they don't have this information it is difficult to see how they can ever advise on change. All they can do is to defend the status quo.

When we consider the complete set of survey activities then listing possible changes appears to be an enormous task. In practice there are some major activities, such as the decision to use one less call-back in an interview survey, or to reduce the sample size by reducing the number of interviews per PSU, or to change the method of editing, which have a major impact on cost and a large potential impact on *L(T)*. By concentrating first on these major areas and by forming the activity changes for various key combinations ΔD, the task can be managed as the case study in Section 4 shows.

The framework proposed is the same as that of Fellegi and Sunter (1973), however, they went a step further and examined the implications of continuous changes. This enabled them to find conditions for an optimal allocation of resources, but these conditions were unrealistic and they concluded that optimal decisions were impractical. Their negative arguments undermined the basic simplicity of the original structure and appear to have discouraged further work. The best is often the enemy of the good, and in my view the systematic framework for discrete changes is workable and could lead to useful improvements in survey quality if adopted.

3.3 The operational decision

The next step is to evaluate the effect on survey errors of the feasible actions represented by the changes ΔD. This feeds into the loss function which we can now denote by *L(T, Δ D)*. The resource allocation problem is solved by choosing the action, ΔD, which minimises *L(T, Δ D)*. Each feasible action affects a subset of the survey errors and the problem facing the operational statistician is the evaluation of the effects. Basically they have to fill in the entries in the matrix of actions errors. For changes in sample size the effects can be computed from knowledge of the structure of sampling errors; for improved coverage due to frame changes the effect is more difficult to compute, but crude estimates should be possible. These estimates can be varied using a sensitivity analysis. The effects of response and measurement errors can be assessed through experiments conducted within the ongoing survey, and this is an area for collaboration between the methodological and operational statisticians. If there are no studies available then the statisticians must use their professional expertise to make subjective judgements about the effects on the errors. Again sensitivity analyses can be employed to allow for a range of subjective judgements.

A source of error that is sometimes overlooked is that due to the timeliness of the publication of the results. All aspects of the survey process take time as well as money. From the point of view of the agency producing the data it is relatively easy to cost time in person-years, and hence to convert time into cash. For the user this is more difficult. How should timeless of data impact on the loss function, *L(T)* ? One approach is to assume that a user at time $t + s$, $s > 0$, employing data collected at time *t*, will make a forecast, $\hat{T}(t + s|t)$, from the data up to and including time *t*, of the actual value of *T* at time $t + s$. The estimated forecasting error can then be added as an extra component to *L(T)*. Typically the forecasting error will increase with *s*, and the more rapid the changes in the variable over time, the larger the forecasting error. Thus this approach should give a reasonable measure of the penalty to be attached to the delay in producing data.

Once the entries in the action by error matrix have been computed, the final step is to evaluate the loss function for each feasible action. The action, or actions, with the smallest value of the loss function can then be considered as a possible action to improve survey quality. If the action with smallest loss is rejected, then the implication is that either the wrong loss function has been chosen, in which case an alternative should be proposed, or that some of the entries in the matrix are in error, in which case they should be identified and changed.

The beauty of the decision framework is that it enables problems to be identified explicitly; it is no longer possible to hide behind general statements about the difficulties of making any changes. Questions about the amount of editing or the effect of non-response, which are the concern of operational statisticians, can now be addressed directly. The search for the best answer will rarely be worthwhile and in most cases approximate answers will provide adequate guidance about the most useful changes.

4. A CASE STUDY

Linacre and Trewin (1993) provides an example of the application of the above principles and is the only published case study that I could find which addresses total survey error in a coherent framework. They consider how various evaluation studies carried out in conjunction with the 1984/5 Construction Industry Survey in Australia could be used to advise on the redesign of the survey for 1988/9. They note that the results can be indicative only, and that subjective assessments must be used when only qualitative information is available. They employ sensitivity analyses when they have doubts about the reliability of their estimates of error. It is worth quoting the introduction to their case study in full:

> *"Evaluation studies are often used to determine effective methods to reduce non-sampling error. A question that frequently arises in practise is how much of the resources available for a collection should be spent on each facet of the collection. How much should be spent on setting up a good quality frame, how much on pilot testing, on field enumeration in preference to mail enumeration, in intensive non-response follow up etc. Each of these 'error reduction' tasks takes resources and the problem is to minimise the total overall error for a collection given one or more fixed resource constraints."*

The paper then considers the operation of the survey and discusses alternative strategies from which they identify a set of possible options for the new survey. Instead of restricting themselves to feasible options they evaluate both the cost and the loss function for each of their options. The loss function chosen is the root mean square error of one key estimator. The options are then plotted on the cost by RMSE graph shown in Figure 1.

The spread of the results is remarkable. Options costing $300k can have the same RMSE as options costing ten times as much. The option originally chosen reduces the cost to one third of the 84/5 cost at the expense of doubting the RMSE. (Note that reducing the sample size to 25% of the 84/5 sample would also double the sampling error). However, another option which reduces costs to one third leads to only a small increase in RMSE, and is approximately twice as effective as the option chosen before the evaluation study.

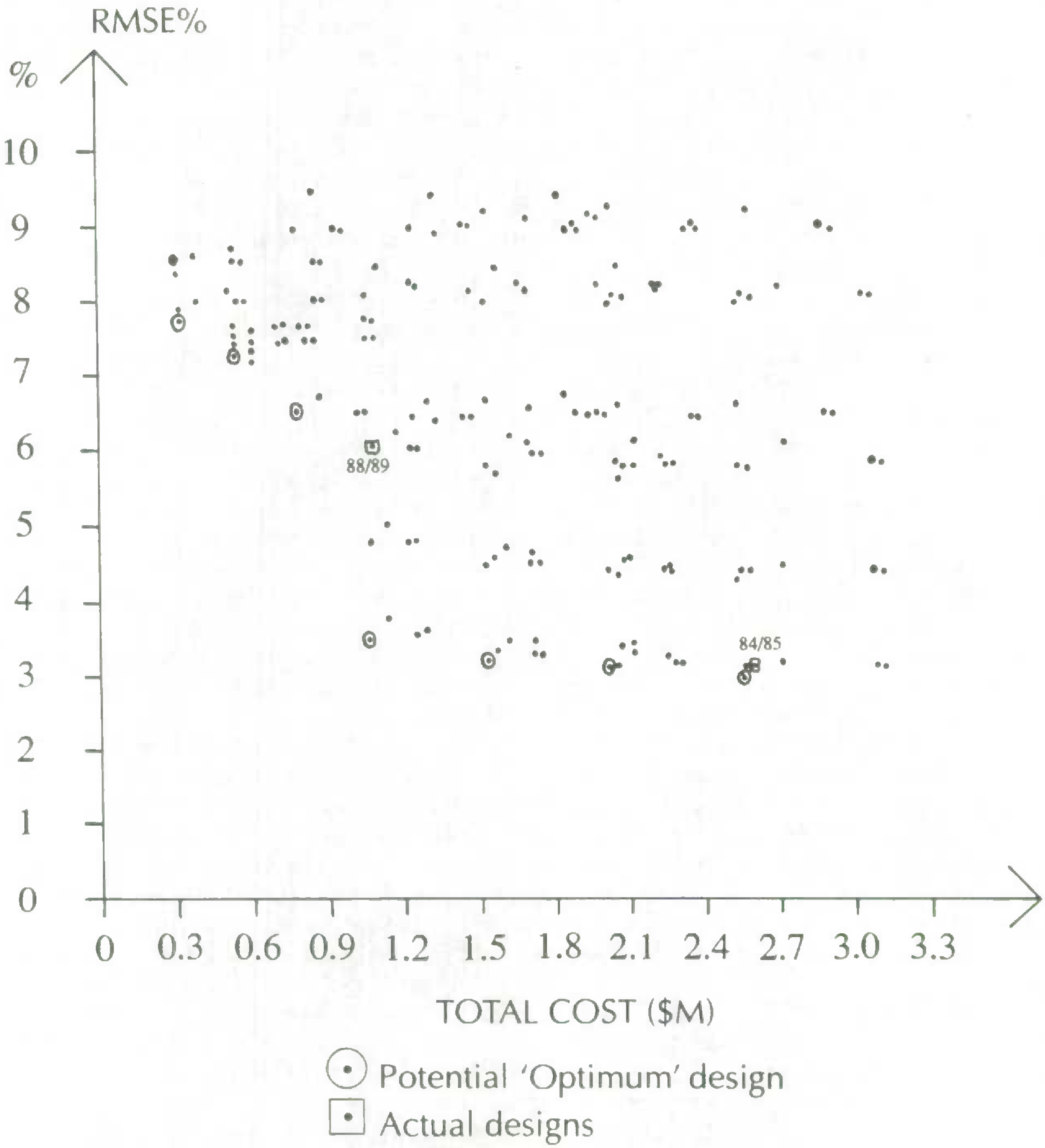


Fig. 1. RMSE% v, total cost for a number of resource allocation options

(Reproduced, with permission, from Linacre and Trewin (1993), Total Survey Design - Application to a Collection of the Construction Industry, *Journal of Official Statistics*, 9, 611-621, Figure 1, p. 618).

This example demonstrates that the systematic structuring of the decision process for the redesign of a survey can lead to dramatic changes in the perceived choice of options. Without this study the first option chosen would have been far from the optimum. The range of options studied could have been considerably reduced by considering only the set of feasible options that met some cost constraints.

5. CONCLUSIONS

The framework suggested by Fellegi and Sunter, as amplified in this paper, provides a mechanism for the systematic consideration of the consequences of alternative options for the allocation of resources over survey operations. Carrying out all stages of the decision process is not easy, but just because something is hard does not mean it should not be attempted. Only by adopting such am approach can those areas be identified where more or better information is required. Statisticians should use their training as statisticians in the management of statistical operations.

It is interesting to conjecture why so little progress has been made to date on the problem of resource allocation. The most obvious explanation is that it is a very difficult problem and that there is tremendous inertia in most systems that mitigates against any change. But surveys are redesigned periodically, and budgets are reduced regularly, so some decisions about change are being made. Why has so little been written about the problem? One issue appears to be that methodologists are frequently used as fire-fighters to fix up analyses after the errors have been perpetrated. In the terms of Groves (1989) they are measurers of errors rather than reducers of errors. Although this is understandable for those working on the secondary analysis of survey data, it also appears to be true for many operational statisticians, and the bulk of the published work on survey errors by official statisticians relates to error estimation in analysis rather than error reduction in design. There also seems to be a culture gap between the operational statisticians, the survey managers, with their concerns about the day to day practical problems of running surveys, and the methodologists, with their interests in statistical theory. The two groups need to come together, not only for survey redesign, but also to improve the routine operations of ongoing surveys. Only when error experiments are imbedded into the routines of all surveys is this likely to happen; and only then will we have the type of information about survey errors that will lead to the desired improvements in survey quality.

The decision theory approach to resource allocation should be particularly useful for handling changes in survey operations brought about by budget changes. Decision theory can also be useful as an instrument for introducing major changes in procedures. The ideas of total quality assurance tell us that the reduction in errors will only come about by changes to the survey system, by moving the problem upstream in the survey process. Arguably the greatest contribution to the improvement in survey operations, with the greatest potential for reducing a range of survey errors, has been the development of computer assisted interviewing. CAPI and CATI have helped to reduce response errors, editing and coding errors, and a host of data processing errors. A qualitative assessment of the impact of the new procedures on both costs and losses should accompany the first introduction of these methods. A systematic statistical analysis can be used to validate the impact of the system changes.

Finally, to quote Linacre and Trewin's conclusion:

> *"If we are to progress further on appropriate allocation of resources to minimise total error, effort must be put into determining and testing these relationships between error and resources for all sources of error. This might be done firstly through identifying other key parameters for these sources of error and then by setting up the models relating these parameters to resource usage. This study has provided a step in this direction. It is hoped that information on any progress made by other statistical organisations in optimal resource allocation for total error will be disseminated amongst the general statistical community, to add to the body of information available in an as yet relatively undeveloped area of considerable importance."*

AKNOWLEDGEMENT

This research was supported by a grant from the Economic and Social Research Council of the UK under its Analysis of Large and Complex Datasets research programme.

REFERENCES

Alexander, C. H. (1994). Discussion of *Sample surveys 1975-1990: an age or reconciliation? International Statistical Review*, 62, 1, 21-28.

Fellegi, I.P., and Sunter, A. B. (1973). Balance between different sources of survey errors - some Canadian experiences, *Proceedings of the 39th Session of the ISI*, 45, 3, 334-355.

Groves, R. M. (1989). *Survey Errors and Survey Costs*, John Wiley and Sons, New York.

Hansen, M.H., Hurwitz, W.N., and Benshad, M.A. (1961). Measurement errors in censuses and surveys, *Proceedings of the 33rd Session of the ISI*, 38, 359-374.

ISI Report (1926). Report on the representative method in statistics, *Bulletin of the ISI*, 22, 1, 359-438.

Jabine, T.B., and Tepping, B. J. (1973). Controlling the quality of occupation and industry data, *Proceedings of the 39th Session of the ISI*, 45, 3, 360-389.

Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.

Linacre, S.J., and Trewin, D.J. (1993). Total survey design - application to a collection of the construction industry, *Journal of Official Statistics*, 9, 3, 611-621

Morgenstern, O. (1963). *On the Accuracy of Economic Observations*, Princeton University Press.

Nathan, G. (1973). Utilization of information on sampling and non-sampling errors for survey design - experiences in Israel, *Proceedings of the 39th Session of the ISI*, 45, 3, 393-406.

Statistics Canada (1987). Statistics Canada's policy on informing users of data quality and methodology, *Journal of Official Statistics*, 3, 83-92.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

INFORMING USERS OF STATISTICS CANADA'S GENERAL SOCIAL SURVEY ABOUT DATA QUALITY

D. G. Paton[1]

ABSTRACT

In order to appropriately interpret statistical information the user needs measures of the quality of that information. When a survey publishes estimates, a quality measure (such as a coefficient of variation or a confidence interval) can be published alongside the estimates. The General Social Survey, however, publishes not only estimates but also a microdata file. Most of the estimates based on GSS data that appear in reports and publications are produced outside of Statistics Canada using the microdata. This paper describes how the General Social Survey provides users of its data with a number of ways to indicate the quality of the estimates they produce and guidelines to follow when publishing those estimates.

KEY WORDS: Data Quality; Design effects.

1. INTRODUCTION

In order to appropriately interpret statistical information the user needs measures of the quality of that information. This paper gives an overview of the way that the General Social Survey (GSS) informs its users about the quality of the data. It begins with a brief description of the GSS and then examines the Statistics Canada policy on the documentation of data quality. It then looks at the specific ways in which the GSS informs users, with some comparisons to other publishers of survey data. Finally, some possible changes to the practices of the GSS are considered.

The General Social Survey

The General Social Survey (GSS) is a household survey that has been conducted annually by Statistics Canada for the last ten years. It has the objective of both gathering data on a broad range of social characteristics such as health, time use, family history, victimization, and education, and of making that data available in aggregate and microdata form for analysis by governments, academics, and other interested organizations and persons.

The target population of the GSS is the non-institutional population of the ten provinces of Canada. The data are collected during telephone interviews using a samples selected with random digit dialling techniques, with supplementary samples of special populations of interest sometimes being selected from list frames. Further detail on the methodology of the GSS can be found in Norris and Paton (1991).

The GSS releases its data in several ways. A selection of the most important and interesting results of the survey are reported in The Daily, Statistics Canada's daily publication announcing the results of surveys and the availability of data. For several of the annual GSS surveys there was a comprehensive Statistics Canada publication providing an in depth look at the results of the survey. However, the most important way in which the GSS data are released is in the form of microdata files.

Statistics Canada Policy

As a Statistics Canada Survey, the GSS releases its data and informs the users of this data in the context of Statistics Canada's policies. Among the various sections in the Policy Manual of Statistics Canada there is one that is about "Informing Users of Data Quality and Methodology" (Statistics Canada, 1992). This policy recognizes that the users of statistical products need some additional information in order to be able to properly work with and interpret the data and that there therefore exists a responsibility to provide that information.

[1] David G. Paton, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The Statistics Canada policy states:

1. Statistics Canada will make available to users indicators of the quality of the data it disseminates and descriptions of the underlying concepts, definitions and methods.
2. Statistical products will be accompanied by or make explicit reference to documentation on quality and methodology.
3. Documentation on quality and methodology will conform to such standards and guidelines as shall from time to time be issued under this Policy.
4. Exemption from the requirements of this policy may be sought in special circumstances using the procedure described below under "Responsibilities".
5. Sponsors of cost recovery surveys and statistical consultation work, for which no data will be disseminated by Statistics Canada, are to be made aware of and encouraged to conform to the applicable elements of the standards and guidelines issued under this Policy.

In the standards and guidelines that accompany this Policy can be seen three broad components to the additional information that are seen as fundamental to the proper interpretation and use of any data:

1) What should the data represent? What are the concepts that are of interest or are of relevance to policies?
2) How did the survey represent these concepts in its data? How different might the data be from the concepts of interest due to the design of the survey?
3) What is the error and uncertainty in the data? How different might the estimates be from some true value by accident?

The first of these components is reflected in the concepts and definitions required under the standards and guidelines, with the specific content including descriptions of the concepts being the survey, the target population, the time frame, and the questionnaire used.

The second of these is largely addressed through the details of the methodology that the standards and guidelines suggest be reported, including the collection method, the sample design, a description of the data capture, processing, editing and imputation, weighting, and estimation.

The need to provide information on the third component, error and uncertainty, should be satisfied by describing the various sources of error and uncertainty and by giving measures of their sizes. Coverage error, sampling error, non-sampling error, and response errors should be considered.

What Does the General Social Survey Do?

The GSS has two principal data releases. The first release takes place in The Daily, and is quite short (the June 6 1995 release consisted of about 2 pages (Statistics Canada, 1995)), with its focus being on providing a few highlights that can be gleaned from the GSS data and to announce the availability of the data. The other principal release is public use microdata file. This file gives data for a large number of variables (400) for each GSS respondent (typically there are about 10,000), and is accompanied by three to four hundred pages of documentation (eg. Statistics Canada, 1994).

In keeping with its small size, the GSS article in The Daily presented little detail about the quality and uncertainty of the data and estimates. A small box of about 30 square centimeters set off a few sentences about the survey's concepts, target population, time frame, collection method, sample size, and response rate. No indication was given that there was uncertainty in the results.

The GSS microdata is the primary GSS product, and it consists of both the file containing the data as well as extensive documentation. The documentation discusses many quality issues. The target population is described in detail and the time frame of the survey is described. The precise questions used are given and references are given to other surveys using the same questions. The collection method, processing methods, editing and imputation are described. The sample design is described and the differences between the sampled population and the target population are discussed. The method by which the estimation weights were derived is described in detail. Mention is made of coverage error, non-response error, and other sources of error.

The documentation does not give precise detail for estimation and variability of estimates because the details of estimation are up to the user of the microdata, and because the variability of an estimate depends on how it is derived. To assist users of the file, the documentation provides guidelines for estimation and for estimating the variability of the resulting numbers.

The guidelines for estimation are of critical importance, since the variability of an estimate is of little importance if the estimate does not reflect the concept of interest. The GSS microdata documentation provides instructions in the use of the data file and the variables and weights included in the file, as well as extensive examples of their use. In the microdata documentation for Cycle 8 (Statistics Canada, 1994) there were 15 pages dedicated to these guidelines and examples.

Ideally, users of the microdata would be able to calculate their own measures of uncertainty for the estimates they make. Unfortunately, the detailed information about the design of the GSS sample that is needed to calculate these measures includes geographic detail that cannot be included in the public use file due to the need to ensure confidentiality. To provide users of the microdata file with alternative methods of estimating the uncertainty of estimates they produce, several guidelines are given. These include design effects, approximate variance tables, minimum sample size guidelines and guidelines for adjusting the weights in some analytical situations. In addition, we offer to calculate, on a cost recovery basis, estimates of the variability of any estimates that the users may be interested in.

Design Effects

While design effects are specific to each estimate it can be useful to have a typical design effect that can be used to adjust variances calculated assuming simple random sampling. The approximate variances for estimates calculated this way will either under- or over-estimate the true variance, depending on whether the true design effect for the estimate is higher or lower than the typical design effect used. Thus the value chosen as the typical design effect for this purpose needs to strike a balance between being too small, in which case many estimates would be considered more precise than they really are, and too large, in which case too many estimates would be considered to have inadequate precision.

To choose a typical design effect the GSS calculates true design effects for a large number of variables from the microdata, typically in excess of 200. The design effect that is published is the 75th percentile of the resulting distribution. This somewhat conservative approach is taken since the precise estimates that will be made by analysts are not predictable. If the 50th percentile were used instead, the published design effects would be 7 to 20% lower, see Table 1.

Approximate Variance Tables

Approximate variance tables provide variance estimates for estimates of totals and proportions that are based on a design effect, the population size, and the sample size. The GSS provides these tables along with instructions for using the tables to estimate the coefficients of variation of ratios, differences, and differences of ratios of estimates, and to calculate confidence limits and t-tests.

Table 1. Design effects based on 50th percentile and 75th percentile, GSS-9.

	50%	75%
CANADA	1.42	1.53
NFLD	1.08	1.27
PEI	0.98	1.25
NS	1.05	1.26
NB	1.19	1.38
QUE	1.17	1.27
ONT	1.11	1.25
MAN	1.06	1.25
SASK	1.03	1.24
ALB	1.14	1.27
BC	1.13	1.26

Minimum Sample Size Guideline

The coefficient of variation for a small proportion estimated from a simple random sample is approximately a simple function of the number of respondents contributing to the numerator of the proportion:

$$cv(p)=\sqrt{fpc}\sqrt{\frac{1}{n*}}.$$

(where n* is the number of respondents contributing to the numerator at the proportion).

This relationship is the basis for a rule of thumb based on generalizing the simple random sample result by applying a design effect, which allows us to produce Table 2.

The microdata documentation recommends that estimates based on fewer than 15 observations not be released or published. This is based on the GSS having a design effect of approximately 1.5 and the use of a maximum allowable cv of 33%.

Variance Inflation Factor

The variance inflation factor can be thought of as the factor by which the variance of the estimate of the total of some variable is inflated due to the variability of the weights when the weights and survey design are unrelated to the variable. This is the factor '1+L' of Kish(1992).

As can be seen from Table 3 this factor will tend to inflate variances calculated from the GSS even more than the use of the design effects would. This helps guard against finding significance unsupported by the data when estimating totals or the parameters in linear regressions, but not for chi-square statistics for contingency tables.

Table 2. Minimum Sample Size by Coefficient of Variation(cv) and Design Effect

C.V.	Design Effect		
(%)	1.0	1.5	2.0
10	100	150	200
14	49	74	98
17	36	54	72
20	25	38	50
25	16	24	32
33	9	15	18

Weight Adjustment

This technique helps with analyses using statistics or computer programmes not well adapted to the use of survey data, especially in those situations where for the algorithms used large weights imply high precision. The technique is to rescale the weights for those cases used in the analysis so that their average is one. This is done by dividing all of the weights by their average. The situations where this is useful have the characteristic that the estimates do not vary when the weights are rescaled, but the variances reported by the analysis programme do. (If the variances do not change, then no harm is done by rescaling.) The common situation is that of the analysis of contingency tables. When the table is a table of proportions rather than counts, then the table does not depend on the scale of the weights.

Directions for Change

In preparing this presentation and reviewing the current practices of the GSS it became apparent that there were some changes that should be made or that should be considered. There are sections in the documentation that need to include more detail. It would also be useful to get some feedback from the users perspective as to what would be useful and where more detail is needed.

the documentation describes how to produce a variety of estimates. This could be expanded to the provision of sample programmes in one or more statistical programming languages, and even to the provision of sample programmes for more complicated tasks such as linear and logistic regressions .currently

Table 3. Variance Inflation Factors(VIF) and Design Effects(deff), GSS-9.

	Deff	
	75%	VIF
CANADA	1.53	1.58
NFLD	1.27	1.28
PEI	1.25	1.28
NS	1.26	1.32
NB	1.38	1.43
QUE	1.27	1.28
ONT	1.25	1.26
MAN	1.25	1.27
SASK	1.24	1.37
ALB	1.27	1.29
BC	1.26	1.28

The minimum sample size is a rule of thumb with a small amount of theoretical justification. Some time should be spent finding out how good a rule of thumb it is and trying to strengthen the theoretical justification.

The possibility of releasing a simplified design should be investigated. If a simplified design could capture most of the sampling variability and could be released without compromising confidentiality, the users themselves could perform meaningful design based analyses. To facilitate this we would need to release sample programmes and recommend analysis tools.

REFERENCES

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8(2), 183-200.

Norris, D. A., and Paton, D. G. (1991). Canada's general social survey: five years of experience. *Survey Methodology*, 17, 227-240.

Statistics Canada (1992). Informing Users of Data Quality and Methodology. Policy (3pp) and Standards and Guidelines (16pp). Uncatalogued, part of Statistics Canada Policy Manual.

Statistics Canada (1994). The 1993 General Social Survey - Cycle 8, Personal Risk, Public use microdata file documentation and user's guide. 429pp. Uncatalogued.

Statistics Canada (1995). General social survey: computers in the workplace. *The Daily*, June 6, 1995, 5-6. Statistics Canada Catalogue no. 11-001E.

SESSION 5

Technical Aspects of Confidentiality

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

GLOBAL RECODINGS AND LOCAL SUPPRESSIONS IN MICRODATA SETS

A.G. de Waal and L.C.R.J. Willenborg[1]

ABSTRACT

Statistics Netherlands applies two techniques to safeguard a microdata set against disclosure, namely local suppression and global recoding. When local suppression is applied some values in some records are replaced by 'missings'. When global recoding is applied some variables are recoded. Ideally, the local suppressions and global recodings should be determined automatically and optimally, i.e. the information loss due to the local suppressions and global recodings should be minimized. In this paper three problems are examined: finding the optimal local suppressions when a microdata set has to be protected by local suppressions only, finding the optimal global recodings when a microdata set has to be protected by global recodings only, and finding the optimal local suppressions and global recodings when a microdata set has to be protected by a mix of both techniques. For the first problem, the socalled local suppression problem, no complicated information measure is required. Several 0-1 integer programming formulations are given depending on the aim of the data protector. For the second and third problem, the so-called global recoding problem and the GR&LS-problem respectively, an elaborate information measure is required, however. In this paper we suggest an information measure based on a suitable entropy measure. Moreover, a verbal description of both the global recoding problem and the GR&LS-problem is presented.

KEY WORDS: Statistical disclosure control; Microdata; Local suppression; Global recoding; Optimization; 0-1 integer programming problems.

1. INTRODUCTION

Two well-known techniques to safeguard a microdata set against disclosure are local suppression and global recoding. When local suppression is applied a number of values of some variables in some records in the microdata set are set to 'missing'. When global recoding is applied a number of variables are recoded. This means that several categories of such a variable are combined to form new categories. For instance, when the categories 'Widowed' and 'Divorced' of the variable 'Marital Status' are combined to form the category 'Widowed or Divorced', then the variable 'Marital Status' is recoded. In the case of global recoding all the records in which these categories occur are affected, whereas in the case of local suppression only some specific records are involved. Local suppression and global recoding, are usually applied in combination, but can also be applied separately. Both techniques aim to produce a data file containing less detailed information, thereby reducing the disclosure risk.

Ideally, we would like to be able to find the global recodings and local suppressions automatically and optimally, i.e. the information loss due to these statistical disclosure control (SDC) measures should be minimum. In this paper we consider three problems, namely the problem **to protect a microdata set** by means of local suppressions only, i.e. the local suppression problem, the problem to protect a microdata set by means of global recodings only, i.e. the global recoding problem, and the problem to protect a microdata set by a combination of global recodings and local suppressions, i.e. the combined global recoding and local suppression problem. This last problem will also be referred to as the

[1] A.G. de Waal and L.C.R.J. Willenborg, Statistics Netherlands, Prinses Beatrixlaan 428, P.O. Box 959, 2270 AZ Voorburg, Netherlands.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

GR&LS-problem in the remainder of this paper.

The local suppression problem does not need a complicated measure for information loss. One can count the number of locally suppressed values. This is a crude but simple way to measure information loss, and one that does not discriminate between the local suppression of values of different variables, which may lead to different information losses. Given this simple measure for the information loss due to local suppressions the local suppression problem can be formulated as an 0-1 integer programming (IP) problem. Of course, in some cases the simple measure for the information loss due to local suppressions may be inappropriate. In such cases a more sophisticated measure of information loss may be called for, such as the one to be discussed shortly which is based on entropy. In Section 2 an introduction to the local suppression problem is given. In particular, a key concept for the local suppression problem, 'minimum unsafe combination', is introduced. This concept allows us to formulate the local suppression problem as an 0-1 IP problem. This is the subject of Section 3. In Section 4 some extensions of the local suppression problem are examined.

In case optimal global recodings are to be determined, the introduction of a measure for information loss seems to be unavoidable. In particular such a measure is needed when the problem is to find the optimal mix of local suppressions and global recodings. A natural choice for such a measure is the entropy. Less straightforward is the choice of a suitable probability model to account for (partial) missingness of information due to global recodings and/or local suppressions. We present a simple model for these situations. An information measure based on the entropy is introduced in Section 5.

The global recoding problem and the GR&LS-problem seem much harder to formalize than the local suppression problem. Hence, these two problems are discussed rather informally. Finding good 0-1 IP formulations for both problems, and solving these, is a task that remains to be done. An introduction to the global recoding problem and the GR&LS problem is presented in Section 6. The 'pure' optimum global recoding problem is discussed in Section 7. The aim is to find the global recodings that eliminate a given set of rare combinations with minimum information loss. Section 8 deals with a description of the GR&LS-problem. Finally, the paper is concluded by a short discussion in Section 9.

This paper is a combination of De Waal and Willenborg (1994), which is devoted entirely to the local suppression problem, and De Waal and Willenborg (1995b), which is devoted to the global recoding problem and the GR&LS-problem.

2. INTRODUCTION TO THE LOCAL SUPPRESSION PROBLEM

SDC rules often describe combinations of categories of identifying variables that have to be checked before a microdata set can be disseminated. Moreover, the rules also describe how many times these combinations have to occur in order to be considered safe for release. In case the frequency of a particular combination is at least a prescribed threshold value then this combination is considered safe. Otherwise the combination is considered unsafe and SDC measures should be applied. In this section as well as in Sections 3 and 4 we assume that these measures consist of replacing certain values by 'missings', i.e. by locally suppressing these values.

The easiest way to determine which values of the variables should be locally suppressed would be to do this for each combination that has to be checked and for each record separately. This can be done in two ways. Firstly, when a value is locally suppressed then this value is set to 'missing' immediately. The resulting microdata is then used to determine whether or not a combination is safe. Secondly, the original microdata set can be used to determine whether or not a combination is safe. However, either way causes problems.

When the first method is used some combinations may incorrectly appear to occur not frequently enough. For example, if we suppress the value 'Baker' in the combination 'Baker'x'Foreigner' appearing in a record then this may have the consequence that later on the combination 'Baker'x'Male' appears to occur not frequently enough. This combination would therefore be considered to be unsafe. However, this combination could occur frequently enough in the original microdata set. So, in fact it should be considered to be safe in that case.

On the other hand, when the second method is used to determine whether a combination is safe and one does this for each record separately then this may also lead to problems. Suppose for instance that the combination 'Baker'x'Foreigner' does not occur frequently enough in the original file and that we decide to locally suppress 'Foreigner'. Suppose furthermore that the combination 'Baker'x'Female' also does not occur frequently enough and in this case we decide to local suppress 'Female'. Then it is not unlikely that we local suppress too much. In case there would be persons who are 'Baker', 'Female' and 'Foreigner' simultaneously then it would have been better if we had local suppressed the category 'Baker' for these persons assuming each of the categories 'Foreigner' and 'Female' occurs frequently enough. The number of local suppressions would have been less if we had local suppressed 'Baker' for these persons.

We can conclude that we cannot decide for each unsafe combination and record separately which values should be suppressed if we want to minimize the number of local suppressions. We have to decide which values have to be locally suppressed for all the unsafe combinations and records simultaneously.

To fix our minds we suppose that it is necessary to check whether certain trivariate combinations of categories of identifying variables occur frequently enough[2]. We start by checking all the univariates. In case a category of a variable is considered safe then we check the bivariate combinations in which this variable occurs. In case a category of a variable does not occur frequently enough, e.g. 'Mayor', then we do not have to check the bivariate combinations involving 'Mayor', e.g. 'Mayor'x'Female'. Then we check the trivariate combinations in which only safe bivariate combinations occur. In case an unsafe bivariate combination occurs in a trivariate combination then this trivariate combination needs not to be checked. For example, if the bivariate combination 'Baker'x'Female' is unsafe then we need not check the trivariate combination 'Baker'x'Female'x'Urk'[3]. After checking the required trivariate combinations we are able to list for all the records the unsafe univariate, bivariate and trivariate combinations. From now on we will call the combinations in this list the *minimum unsafe combinations*.

A consequence of the above approach is that whenever we locally suppress a value in a minimum unsafe n-variate combination then the resulting (n-l)-variate combination will be safe. This property of the minimum unsafe combinations makes it easy to find the minimum number of local suppressions.

As a concluding remark to this section we would like to point out that it is not essential that the threshold value to determine whether a combination is safe or not is a fixed number. In fact, the threshold value may depend on the combination. For instance, the threshold value of a univariate 'combination' may be higher than the threshold value of a bivariate combination. In this case it may happen that a bivariate combination 'Mayor'x'Female' is considered safe while 'Mayor' is considered unsafe when examined univariately. When this happens 'Mayor' is still considered to be a minimum unsafe combination. The 0-1 IP formulations of Section 3 and 4 remain valid for such a case.

3. MINIMIZATION OF THE NUMBER OF LOCAL SUPPRESSIONS

The first problem we consider in this paper is the problem of finding a minimum number of local suppressions such that the resulting microdata set is considered safe. This problem can be formalized as follows. Suppose we need to suppress some categories of variables in some records. For each category j of a minimum unsafe combination in record i we introduce a dummy variable Y_{ij}. This dummy variable is equal to 0 if category j in record i is not suppressed or if category j does not occur in record i, otherwise it is equal to 1. For each minimum unsafe combination and for each record we have the constraint stating that at least one category of a minimum unsafe combination in a record must be suppressed. In other words, the sum of the Y_{ij}'s of the corresponding categories j is at least 1. As we have remarked before this constraint is necessary and sufficient in order to make this combination safe. As a target function we use a weighted sum of the Y_{ij}'s.

In mathematical terms we consider the following 0-1 IP problem. Let the total number of unsafe records be denoted by I and the total number of categories of the unsafe combinations by J. After renumbering the records and the variables the dummy variables $Y_{ij}(i=1,...,I; j=1,...,J)$ must satisfy

$$y_{ij} = \begin{cases} 1 & \text{if category } j \text{ in record } i \text{ is suppressed} \\ 0 & \text{if category } j \text{ in record } i \text{ is not suppressed,} \\ & \text{or if category } j \text{ does not occur in record } i \end{cases} \qquad \textbf{(1)}$$

Suppose there are K minimum unsafe combinations in the microdata set. Let c_{jk} be equal to 1 if category j occurs in minimum unsafe combination $k(j=1,...,J; k=1,...,K)$ and 0 otherwise. The constraints of the problem are given by

[2] We only consider the identifying variables, because our measures to protect a microdata set involve such variables only. Whenever we refer to (a category of) a variable we will mean (a category of) an identifying variable.

[3] Urk is a small picturesque fishing - village in the Netherlands.

$$\sum_{j=1}^{J} c_{jk}y_{ij} \geq d_{ik}, \text{ for all } i=1,\ldots,I; k=1,\ldots,K, \qquad (2)$$

where d_{ik} equals 1 if minimum unsafe combination k occurs in record I, otherwise d_{ik} equals 0. The constraints given by (2) must hold because at least one category in each minimum unsafe combination has to be suppressed.

We consider the following target function

$$\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}y_{ij}, \qquad (3)$$

where w_{ij} denotes the nonnegative weight of category j in record I which needs to be specified by the user. Our problem is to minimize target function (3) under the constraints given in (2).

Note that if we choose all the weights w_{ij} equal to one then the aim is to minimize the number of local suppressions. Because the weights in target function (3) may be arbitrary nonnegative numbers the problem stated above is more general. The weights allow one to differentiate between the relative importance of specific categories in specific records as far as local suppression is concerned.

The above problem can be solved by using a standard algorithm to solve 0-1 IP problems, such as a branch-and-bound algorithm (Cf. Nemhauser and Wolsey, 1988). Moreover, the problem can be reduced to a number of smaller problems. First of all, it can be decomposed into subproblems for each record separately. For each record I target function (3) has to be replaced by target function

$$\sum_{j=1}^{J} w_{ij}y_{ij}, \qquad (4)$$

The constraints to be considered for this problem consist of all those given in (2) as far as they pertain to record I. Even this subproblem for each record can sometimes be partitioned into a number of smaller subproblems. Consider the minimum unsafe combinations of a particular record to be the vertices of a graph. If two minimum unsafe combinations have a category in common, then they are joined by an edge. This graph may be disconnected. In that case it consists of several connected subgraphs that are mutually disconnected. Each subgraph corresponds to a subproblem, namely the problem of minimizing (4), under the constraints that the minimum unsafe combinations corresponding to the vertices are made safe. So, sometimes we will be able to reduce the original problem to a number of smaller subproblems. But even these subproblems may sometimes be reduced to still smaller problems, some of which may be trivial. This further reduction follows from the observation that only the dummy variables corresponding to categories that occur in more than one minimum unsafe combination have to be considered. Combinations that are still unsafe after some of these categories have been suppressed (in an optimal way) can be made safe according to a list of priorities. No optimization problem has to be solved for these remaining unsafe combinations because they do not have a category in common.

We give some examples to make this observation somewhat clearer. Suppose we want to minimize the number of local suppressions in a specific record. In case none of the minimum unsafe combinations have a category in common, there is no optimization problem. For each minimum unsafe combination in such a record a single value appearing in this combination has to be suppressed, which can be chosen arbitrarily. In practice the values that are to be suppressed could be determined by means of a list of priorities. This list of priorities have to be supplied by the user. An example of such a list of priorities is the following: first suppress a category of the variable 'Residence', then of the variable 'Sex', then of the variable 'Nationality' and finally of the variable 'Occupation'. If in a record two unsafe combinations occur, 'Baker'x'Foreigner' and 'Female'x'Urk', then 'Urk' and 'Foreigner' will be suppressed in this record.

In case some of the minimum unsafe combinations do have a common category then the situation is somewhat more complicated. Suppose that in a record the combinations 'Baker'x'Female' and 'Baker'x'Foreigner'x'Urk' are minimum unsafe ones. In this case we can minimize the number of local suppressions by suppressing 'Baker'. Both resulting combinations, i.e. 'Female' and 'Foreigner'x'Urk', will be safe then. Note that we do not use a list of priorities for this record. However, for records in which minimum unsafe combinations occur with several categories in common this might be necessary.

So, in practice we can expect that the general optimization problem will be reduced to a number of small subproblems. This implies that it may even be feasible to try all possibilities in order to minimize the target function. However, a standard algorithm to solve 0-1 IP problems may be faster.

Throughout Section 3, 4 and 5 we illustrate the solutions to the problems by means of a standard example. In this example there are eleven unsafe records and 21 different categories. These records contain the following minimum unsafe combinations:

record 1: 'A'x'B'and'B'x'C'
record 2: 'A'x'D'and'A'x'E'
record 3:'C'x'F'
record 4: 'G'x'H'and'H'x'I'
record 5:'J'x'K'
record 6:'J'x'L'
record 7: 'M'x'N'and'N'x'O'
record 8: 'M'x'O'
record 9: 'P'x'Q' and 'Q'x'R'
record 10:'S'x'T'
record 11: 'S'x'U'

Example: If target function (4) has weights w all equal to one, then a solution of the problem considered in this section is given by:

suppress in record 1: 'B'
suppress in record 2: 'A'
suppress in record 3: 'F'
suppress in record 4: 'H'
suppress in record 5: 'J'
suppress in record 6: 'J'
suppress in record 7: 'N'
suppress in record 8: 'O'
suppress in record 9: 'Q'
suppress in record 10: 'T'
suppress in record 11: 'U'

So, 11 values are locally suppressed and 10 different categories are involved.

4. EXTENSIONS OF THE LOCAL SUPPRESSION PROBLEM

In this section we discuss a number of problems that are similar to me problem discussed in Section 3. First of all, instead of minimizing the total number of local suppressions the user of the data could want to minimize the number of different categories that are suppressed. A reason for this could be that he considers a category that is locally suppressed in some records to be unsuited, or hardly suited, for statistical analysis. In other words, locally suppressed categories are of no, or only limited, value to him.

Minimization of the number of different locally suppressed categories:

We can formulate this second problem as follows. First we introduce some new dummy variables. For each category *j* that occurs in a minimum unsafe combination we introduce a dummy variable z_j, defined as

$$z_j = \begin{cases} 1 & \text{if category } j \text{ is locally suppressed} \\ 0 & \text{if category } j \text{ in not locally suppressed.} \end{cases} \tag{5}$$

Note that the z_j's are independent from the records. The following constraints have to be satisfied.

$$\sum_{j=1}^{J} c_{jk} z_j \geq 1, \text{ for all } k=1,\ldots,k. \tag{6}$$

We consider the following target function.

$$\sum_{j=1}^{J} z_j. \tag{7}$$

Target function (7) must be minimized under the constraints given by (6). The optimization problem that then arises is a (minimum cardinality) set-covering problem.

This problem is harder to solve than the problem of Section 3, because here the records cannot be considered independently. However, in many cases the problem can be reduced to smaller subproblems, because the remarks made at the end of Section 3 apply to this case as well. The problem can sometimes be further reduced to subproblems corresponding to connected subgraphs. In this case the minimum unsafe combinations correspond to the vertices of a graph. Two vertices are joined by an edge if the corresponding minimum unsafe combinations have a category in common and both combinations occur at least once in a record simultaneously. Furthermore only the dummy variables corresponding to categories that occur in more than one minimum unsafe combination have to be optimized. Some algorithms to solve the above problem have been examined in Van Gelderen (1995). It appears that a near optimal solution can generally be found within a short period of time.

When this problem has been solved the number of different categories that are locally suppressed is minimized. However, for some records more categories than necessary may have been locally suppressed. In order to overcome this problem each record has to be checked separately. If too many categories of minimum unsafe combinations in this record have been locally suppressed, i.e. more than the number of minimum unsafe combinations, then some of these local suppressions may be replaced by their original value. Which locally suppressed categories are to be replaced by their original values can be determined by means of the list of priorities and the constraint that the record should remain safe.

This problem can be extended by replacing (7) by a weighted sum of the z_j's. This would enable the user to indicate the importance of each category. Important categories should be given a high weight, unimportant ones a low weight. The resulting problem can sometimes be decomposed into a number of subproblems in a similar way as described in Section 3.

Example: (continued) We consider our example again. A solution of the problem considered in this section is given by:

suppress in record 1: 'A' and 'C'
suppress in record 2: 'A'
suppress in record 3: 'C'
suppress in record 4: 'H'
suppress in record 5: 'J'
suppress in record 6: 'J'
suppress in record 7: 'M' and 'O'
suppress in record 8: 'O'
suppress in record 9: 'Q'
suppress in record 10: 'S'
suppress in record 11: 'S'

So, 13 values are locally suppressed and 8 different categories are involved.

Maximization of the number of different locally suppressed categories given that the number of local suppressions has been minimized:

The problems discussed so far can be extended a bit. Suppose that the number of local suppressions has been minimized by solving the 0-1 IP problem of Section 3. Suppose furthermore that among these solutions we want to find the solution that suppresses a maximum number of different categories. As a result the local suppressions will probably spread more or less evenly over the various categories.

This problem can be formalized as follows. Let the minimum number of local suppressions be denoted by N_{min}. This number is known because we assume that the problem of Section 3 has been solved. We want to use both the variables y_{ij} and the variables z_j in one problem. The variable z_j should be equal to one if and only if there is a y_{ij} equal to one for some I. This can be achieved by using a large number W and introducing the following relations.

$$Wz_j \geq \sum_{i=1}^{I} y_{ij}, \text{ for all } j=1,..., J \tag{8}$$

and

$$y_{ij} \geq z_j, \text{ for all } i=1,...,I;\ j=1,..., J. \tag{9}$$

As we want the minimum number of local suppressions we have to add the following constraint.

$$\sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} = N_{min}. \tag{10}$$

The target function we consider is given by (7). This target function must be maximized under the constraints given by (2), (8), (9) and (10).

Example: (continued) We consider our example again. A solution of the above problem is given by:

suppress in record 1: 'B'
suppress in record 2: 'A'
suppress in record 3: 'F'
suppress in record 4: 'H'
suppress in record 5: 'K'
suppress in record 6: 'L'
suppress in record 7: 'N'
suppress in record 8: 'M'
suppress in record 9: 'Q'
suppress in record 10: 'T'
suppress in record 11: 'U'

So, 11 values are locally suppressed and 11 different categories are involved.

Minimization of the number of different locally suppressed categories given that the number of local suppressions has been minimized:

Similar to the above problem the user of the data could want to suppress as few different categories as possible while at the same time suppressing as few values as possible. In this case the target function (7) must be minimized under the constraints given by (2), (8), (9) and (10).

Example: (continued) We consider our example again. A solution of the above problem is given by:

suppress in record 1: 'B'
suppress in record 2: 'A'
suppress in record 3: 'F'
suppress in record 4: 'H'
suppress in record 5: 'J'
suppress in record 6: 'J'
suppress in record 7: 'N'
suppress in record 8: 'M'
suppress in record 9: 'Q'
suppress in record 10: 'S'
suppress in record 11: 'S'

So, 11 values are locally suppressed and 9 different categories are involved.

Minimization of tire number of local suppressions given that the number of different locally suppressed categories has been minimized:

The final problem we consider is the following. Suppose that the number of different locally suppressed categories has been minimized by solving the first 0-1 IP problem considered in this section. Suppose furthermore that among these solutions we want to find the solution that suppresses a minimum number of values.

Let the minimum number of different locally suppressed categories be denoted by $M_{\min}$. This number is known because we assume that the corresponding problem has been solved (to good approximation). We introduce the following constraint.

$$\sum_{j=1}^{J} z_j = M_{min} \ . \qquad (11)$$

In this case we have to minimize (3) with all w_{ij}'s equal to one under the constraints given by (2), (8), (9) and (10).

Example (continued): For the last time we consider our example. A solution of the above problem is given by:

suppress in record	1: 'A'and 'C'
suppress in record	2: 'A'
suppress in record	3: 'C'
suppress in record	4: 'H'
suppress in record	5: 'J'
suppress in record	6: 'J'
suppress in record	7: 'N'
suppress in record	8: 'M'
suppress in record	9: 'Q'
suppress in record	10: 'S'
suppress in record	11: 'S'

So, 12 values are locally suppressed and 8 different categories are involved.

The local suppression problems that are examined in this paper do not take dependencies between variables into consideration. For example, suppressing the value 'Female' of the variable 'Sex' is useless when the value of the variable 'When did give birth to your first child?' is '1976'. The value of the latter variable implies that the value to the former one is 'Female'. The local suppression problems considered in this paper can be extended to take such dependencies into account (Cf. De Waal and Willenborg, 1995a). We do not consider these extensions here, instead we go on with the global recoding problem and the GR&LS-problem.

5. PRELIMINARY REMARKS ON THE GLOBAL RECODING PROBLEM

We now turn to the global recoding problem and the GR&LS-problem. Like in the previous sections, we assume that protecting a microdata set amounts to making sure that certain combinations of categories of identifying variables occur frequently enough, that is, more frequently than a certain threshold value. When a certain combination of categories of identifying variables does not occur frequently enough, then we will either locally suppress one or more values of this combination of categories, or we will globally recode one or more variables in such a way that the corresponding combination of the globally recoded categories occurs frequently enough.

Both global recoding and local suppression lead to loss of information. If we want to offer the users of the data as much information as we can, we have to determine the global recodings and the local suppressions in such a way that they transform an unsafe microdata set into a safe one while retaining as much information as possible. To do this a number of problems have to be solved: a suitable information measure has to be found, the global recoding problem and the GR&LS-problem have to be formulated mathematically and, finally, algorithms to solve these problems (to good approximation) have to be developed.

Before we present these descriptions of the global recoding problem and the GR&LS problem the precise nature of the global recodings should be clarified. In order to perform such global recodings for a variable, a proximity structure should be defined on the domain of this variable. To fix our minds suppose that on a domain D of some (key) variable a metric, or distance function, d has been defined. This metric defines the distance between any two categories C_1 and C_2 of D, expressed as $d(C_1\ ,\ C_2)$. We shall assume that d can also take the value '∞' (infinite), in order to express that two categories are too far apart to ever consider them to be combined into a single category. Now, let there be a category C_1 which has to be combined with one or more other categories of D in a global recoding action. Then it is clear that one should look for categories 'in the neighborhood' of C_1 , which in our case is supposed to mean that $d(C_1\ ,\ C_2) < \tau$ for a certain critical value τ. Suppose that there is another category C_3 that is sufficiently close to C_1, i.e. $d(C_1\ ,\ C_3) < \tau$. In order to accept $C_1 + C_2 + C_3$ as a valid global recoding it should also hold that $d(C_2\ ,\ C_3) < \tau$. We can continue this process of determining valid global recodings until all valid global recodings involving C_1 have been determined. At any step it should be checked that the mutual distances of the categories to be recoded into one category are less than τ.

A matter that needs further investigation is how to define a proximity structure on the domain of a variable that can be used for global recoding purposes. The following possibilities seem to be convenient ways.

1. By actually deriving a metric on D through the application of suitable (multivariate) statistical techniques, such as used in clustering. An example of such a technique is the following. Let the variable under consideration be denoted as V. Let C_1 and C_2 be two categories of V. Select a variable V' that is highly correlated with V. Determine the frequencies of the contingency table $V \times V'$. When the frequencies of the row defined by $V=C_1$ are given by $(n_1, \ldots, n_s)$ and the corresponding frequencies of the row defined by $V=C_2$ by $(m_1, \ldots, m_s)$, then the distance between C_1 and C_2 can be defined by

$$d(C_1,C_2) = \sum_{i=1}^{s} \frac{(n_i - m_i)^2}{(n_i + m_i)}, \qquad (12)$$

   where $(n_i - m_i)^2 / (n_i + m_i)$ equals zero by definition when both n_i and m_i are zero. A number of categories may be collapsed into a single one when the distance between each pair is less than τ.

2. By using an elementary graph on D which defines directly neighboring points. The proximity structure to be used for global recoding can be derived from this graph for a given threshold τ in a similar way as in the previous point, by assuming that each edge has length 1. For some kinds of variables a 'direct neighboring' graph suggests itself. For instance, for ordinal variables the ordering of the categories can be used, for hierarchical variables the hierarchy (which is a tree when viewed as a graph), and for geographical areas the contiguity of such areas.

So there are several ways to determine set proximity structures for a key variable. This may be a rather laborious process initially, but once such an exercise has been carried out for a key variable the result can also be used for other microdata sets in which this variable appears.

Instead of determining a proximity structure for each key variable a data protector could also specify the valid global recodings himself. The problem is then to select the valid global recoding (and local suppressions) that minimize the information loss while protecting the microdata set. Because the number of valid global recodings specified by a data protector will generally be less than when proximity structures are used the complexity of the global recoding and the GR&LS-problem will be reduced.

6. MEASURES FOR INFORMATION LOSS

In order to be able to compare the information content of a microdata set before and after global recodings and local suppressions we need to construct a suitable information measure. We suggest to use the entropy to measure the loss of information due to global recodings and local suppressions. The entropy is a well-known measure for the uncertainty about the value of a variable. The higher the uncertainty about the actual value of a variable, the higher the information loss. We begin by explaining how the information loss due to global recodings and local suppressions can be measured for each variable and record separately.

Suppose, for instance, that the variable 'Marital Status' can assume four values, 'Widowed', 'Divorced', 'Married' and 'Unmarried'. Suppose, furthermore, that two categories, 'Widowed' and 'Divorced' of this variable have been combined into a single category, 'Widowed or Divorced'. In this case a user does not know the original value of the variable when he comes across this combined category in a record. He has to guess whether the original value is either 'Widowed' or 'Divorced'. So, for the user there is some uncertainty about the original value of 'Marital Status' in this case. When, on the other hand, the value of 'Marital Status' is locally suppressed, then the user has to guess which of the four possible categories is the correct one. Again, there is some uncertainty about the original value of 'Marital Status' for the user. In both cases, global recoding and local suppression, the uncertainty about the original value of 'Marital Status' can be measured by means of the entropy.

In case 'Widowed' and 'Divorced' have been combined into 'Widowed or Divorced' then the entropy is defined in the following way. Suppose the probability that the original value of Marital Status' is 'Widowed' equals p_w and the probability that it is 'Divorced' equals p_D The entropy H, i.e. the information loss due to global recoding, is given by

$$H = -p_w' \log(p_w') - p_D' \log(p_D'), \qquad (13)$$

where p_w' and p_D' are the conditional probabilities that the original value of 'Marital Status' equals 'Widowed' and 'Divorced', respectively, given that the recoded value of Marital Status' equals 'Widowed or Divorced'. In mathematical terms:

$$p_W' = \frac{p_W}{p_W + p_D} \text{ and } p_D' = \frac{p_D}{p_W + p_D} . \qquad \textbf{(14)}$$

In case 'Marital Status' is locally suppressed then the entropy is defined in the following way. Suppose that the probability that the original value of 'Marital Status' is 'Widowed' equals p_W, the probability that it is 'Divorced' equals p_D, the probability that it is 'Married' equals p_M and the probability that it is 'Unmarried' equals $p_U = 1 - p_W - p_D - p_M$. The entropy H, i.e. the information loss due to local suppression, is given by

$$\begin{aligned} H = -p_w \log(p_w) - p_D \log(p_D) \\ -p_M \log(p_M) - p_U \log(p_U) \end{aligned} \qquad \textbf{(15)}$$

Note that local suppression is in fact an extreme case of global recoding: it is a recoding of all categories into a single one. This property is incorporated in the above entropy measure. An important difference between global recoding and local suppression is however that global recoding leads to an information loss in all the records that contain at least one of the recoded values while local suppression leads to an information loss in the corresponding record only.

In general, when categories C_1 C_2 ,..., C_n of a variable V are combined into a single one, $C_1 + C_2 + ... + C_n$, then the information loss due to this global recoding is given by

$$H = -\sum_{i=1}^{m} p_i' \log(p_i') \ , \qquad \textbf{(16)}$$

where p_i' is the conditional probability that the original value of V is equal to C_i given that the recoded value equals $C_1 + C_2 + ... + C_n$, and m is the number of categories of V. Note that $p_j' = 0$ when C is not part of the recoded category $C_1 + C_2 + ... + C_n$. When $p_j' = 0$ then $p_j' \log(p_j')$ equals zero by definition. When the value of V is locally suppressed, then the information loss is also given by (16).

The information loss for a specific variable and a specific record due to global recoding and local suppression can be evaluated by means of the same formula (16). There is, however, a fundamental problem that has to be solved in order to apply formula (16) in practice: the probabilities p_i cannot be computed without assuming a particular model for the user of the data. Several such models can be assumed. For instance, the user of the data can exploit all the information contained in the records to determine the probabilities p_i. Many, often rather elaborate, multivariate techniques can be applied to do this. One can also assume that a user will take into consideration that some 'missings' are a consequence of SDC. When a 'missing' is caused by SDC then the original value occurs rarely in combination with some other categories. The user may apply this knowledge to evaluate the probabilities p_i. However, we will assume that the probabilities p_i are calculated in a rather straightforward way, to be explained below.

Suppose that a variable V takes m possible values. The frequency of the I-th category, C_i, of variable V in the population is equal to N_i. Let the number of individuals in the target population of variable V be given by N_V. Note that this number N_V may differ for different variables. For instance, the question, and corresponding answer, "What is your age?" has quite a different target population than the question "When did you give birth to your first child?". The former question refers to the entire population whereas the latter question refers only to women who gave birth to a child. The (unconditional) probability distribution of V is used to evaluate the probabilities p_i, i.e. p_i is given by

$$p_i = \frac{N_i}{N_V} . \qquad \textbf{(17)}$$

In practice one often does not know N_V and the N_i's while in most cases only the scores of a sample of the population are known. We can solve the problem in the following way. Let the frequency of category C_i of variable V in the sample be denoted by n_i and the number of individuals of the target population of V in the sample by n_V. With respect to the 'missings' we make the following assumptions. Firstly, a 'missing' can be caused by an individual who is not a member of the target population of variable V. For such an individual the value of variable V is 'missing' by definition. Secondly, we assume that the probability that a 'missing' is caused by an individual in the target population is equal for all individuals in the target population. Under these assumptions we can estimate the probability p_i given by (17) by

$$\hat{p}_i = \frac{n_i}{n_V} . \qquad \textbf{(18)}$$

Since (17) can often not be applied, we use (18) to estimate the entropy of a global recoding or local suppression.

The estimate given by (18) for the probability given by (17) is a rather unsophisticated one. When (some of) the numbers n_i are small the estimates for the corresponding probabilities can be quite bad. More sophisticated methods to estimate the probabilities given by (17) can be proposed.

The probabilities p_i need to be evaluated in order to compute the information loss due to global recodings and local suppressions. Instead of the above, rather simple, model to evaluate these probabilities p_i other models can be developed. We would like to point out, however, that it is not necessary to use a very elaborate model, because no formal information measure can fully quantify the 'true' information loss in a microdata set due to global recoding and local suppression. In fact, the information loss in a microdata set due to global recoding and local suppression is for a substantial part determined by subjective considerations by the users of this data set. Moreover, the information measure is only meant as a tool to find good global recodings and local suppressions; finding the 'best' global recodings and local suppressions is impossible because 'best' is defined in subjective terms.

So far we have concentrated on the information loss due to globally recoding a variable or locally suppressing a value of a variable in a record. However, we need to measure the information loss in the entire microdata set due to global recodings and local suppressions. The measure for the information loss in an entire microdata set that we propose to use is a weighted sum of the information losses for the variables in the records. When the information loss of the *I*-th variable in the *j*-th record is given by H_{ij}, then the information loss in the entire microdata set is given by

$$\Sigma_i \, \Sigma_j \, w_{ij} \, H_{ij}, \qquad \textbf{(19)}$$

where w_{ij} is a nonnegative weight and the sum is taken over all variables *I* and records *j* . The larger the weight w_{ij} the more important the value of the *I*-th variable in the *j*-th record. The weights can be chosen by a user on subjective grounds. The information loss H_{ij} of *I*-th variable in the *j*-th record due to global recoding or local suppression is measured in the same way as described in this section.

To keep the situation as simple as possible we suggest to make the weights equal for different records, i.e. we suggest to replace (19) by

$$\Sigma_i \, \Sigma_j \, w_i \, H_{ij}. \qquad \textbf{(20)}$$

In other words, weights are permitted to differ only for the variables and not for the records. H_{ij} itself is measured by means of (16). Moreover, we suggest to make most w_i's equal to one. Only for those variables that are clearly more important than others we suggest to use a higher weight. Likewise, only variables that are obviously less important than others should have a weight less than one.

Alternatively, one can make the weights equal for different variables, i.e. (19) can be replaced by

$$\Sigma_i \, \Sigma_j \, w_j \, H_{ij}. \qquad \textbf{(21)}$$

In this case the sampling weights of the records are natural choices for the weights w_j.

7. THE GLOBAL RECODING PROBLEM

The optimum global recoding problem can be stated in general terms. For a given set of minimum unsafe combinations the aim is to apply global recodings of (some of) the variables involved in these combinations in order to eliminate them, with in the end a minimum information loss incurred to the microdata set. This is all very easily stated, but finding such global recodings seems to be a difficult task. Contrary to the local suppression problem, where setting at least one value in a minimum unsafe combination to missing would automatically produce a safe combination, such a 'luxury' is not present at the global recoding problem: application of a valid global recoding involving two categories, of which one appears in an unsafe combination, does not guarantee that this unsafe combination is 'eliminated', i.e. absorbed into a new one that occurs frequently enough.

For each valid global recoding one has to find in the data set which records it would affect if it would be carried out, and whether it would be effective in eliminating a rare combination. If so, the global recoding is potentially of interest and the information loss that would result from application of the corresponding global recoding can be calculated.

The global recoding problem is closely related to a problem occurring when protecting a set of linked tables, i.e. tables with common variables obtained from the same base file. Suppose one wants to use the same categorization for each variables in each table where this variable occurs. Suppose furthermore that we want to protect the tables against disclosure by means of recoding the variables only. In this case the tables can be compared to the unsafe combinations in the case of microdata. The only difference is the underlying criterion to determine whether or not a table or combination, respectively, is considered unsafe.

8. THE GR&LS-PROBLEM

For a given set of minimum unsafe combinations the aim of the GR&LS-problem is to recode (some of) the variables globally and to suppress some values locally such that the resulting microdata set is safe and the information loss due to these global recodings and local suppressions is minimal. Solving this GR&LS-problem, or the global recoding problem, optimally, i.e. in such a way that the total loss of information is minimal, is a hard optimization problem.

When the global recoding problem would be solved, the GR&LS-problem could be solved, in theory, by considering all partitionings of the set of unsafe combinations. Given a set U of unsafe combinations, consider all partitions of U into two sets U_1 and U_2. The general idea is that the combinations in U_1 are eliminated by optimal global recoding and those in U_2 by optimal local suppression. To be a bit more precise, we proceed as follows. First we eliminate all combinations in U_1 by optimal global recodings. Then we determine for the new key variables the set of unsafe combinations that are left (in extreme cases this set may be empty). Let this set be denoted by $U_2^{\prime}$. Now eliminate the combinations in this set through the application of optimum local suppressions. The information loss for each suppressed value is assumed to be measured by the entropy.

The information loss due to either the global recoding or local suppression can be calculated. The total information loss is the sum of these partial information losses. In order to finish considering this partition $\{U_1, U_2\}$ reverse the roles of U_1 and U_2 in the procedure just described and repeat it for the new situation.

The number of bi-partitions $\{U_1, U_2\}$ of U to be considered can be quite big: if there are n rare combinations, then 2^n bi-partitions are possible. Some clever method should be devised that efficiently searches through the set of bi-partitions for a given set of minimum unsafe combinations.

9. DISCUSSION

Automated optimal local suppression of categories of records in order to safeguard these records against statistical disclosure seems not too hard to implement. This is especially true when the total number of suppressed categories is to be minimized, because the resulting problems, one problem for each unsafe record, are all very small. In case the number of different locally suppressed categories is to be minimized the problem is somewhat harder. However, in many cases the problem can be decomposed into a number of smaller, and therefore easier to solve, problems. So, it seems possible to solve this problem fairly efficiently as well. Finally, it is also possible to solve some extensions of these two basic problems such as minimizing the number of different locally suppressed categories given that the number of locally suppressed values is minimal.

The basic approach to the problem of determining the local suppressions as outlined in this paper, i.e. minimizing the number of local suppressions by means of solving a 0-1 IP problem, can be useful for similar problems as well. For example, the user might wish to replace some categories in an unsafe combination by other categories rather than locally suppress these values. This problem can be solved, as far as the optimization problem is concerned, in a similar way as the problem of Section 3. We start by locally suppressing values in minimum unsafe combinations in an optimal way. Then we have to impute values for the suppressed ones. For this we have to determine the possible imputations for each record. An imputation is possible in case the resulting record is safe, i.e. if all the combinations do not occur in the list of minimum unsafe combinations, and all specified edit rules (if any) are satisfied. The main problem here is not the optimization problem, but rather the imputation problem to ensure the integrity of the data. The formulation given here shows that the local suppression problem is related to the edit and imputation problem, i.e. the problem of localizing errors made by respondents in a questionnaire and replacing them by better answers. Whether or not an error has been made by a respondent is checked by means of edits, i.e. rules that describe constraints that must be satisfied by the data in a record. The edits can be compared to the frequency checks that have to be made in our SDC procedure. In fact, the frequency checks can be considered as macro-edits.

The possibility of automated optimal local suppression of categories by means of a computer program such as ARGUS (Cf. Pieters and De Waal, 1995; De Waal and Pieters, 1995) highly speeds up the process to produce a safe microdata set. Moreover, it allows the potential users of a microdata set to participate in the production of a safe version of this set which is most suitable for their purposes, given the constraints due to the SDC-rules. Both the statistical office and the users of a microdata set can benefit from this. The statistical office because the users of a microdata set can be less critical of the disclosure control procedure; the users because they can obtain the most useful information from the data rather quickly under the given restrictions.

Automated global recoding of variables to protect a microdata set against disclosure is much harder to achieve. This is true both for the case where a microdata set has to be protected by global recodings only and the case where a microdata set has to be protected by a combination of local suppressions and global recodings. One of the problems is that a measure has to be constructed to evaluate the information loss due to local suppressions and global recodings. In this paper we suggest a measure based on the entropy.

Formulating the global recoding problem and/or the GR&LS-problem as an optimization problem seems to be rather difficult. In this paper we have limited ourselves to a verbal description of these problems. Finding a good formulation of the global recoding problem and the GR&LS-problem is a problem that remains to be solved. A practical solution to overcome part of the problems is to let the data protector specify the valid global recodings himself. In this way the complexity of the global recoding problem and the GR&LS-problem can be considerably reduced.

REFERENCES

De Waal, A.G., and Pieters, A.J. (1995). ARGUS User's Guide, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., and Willenborg, L.C.R.J. (1994). Minimizing the Number of Local Suppressions, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., and Willenborg, L.C.R.J. (1995a). Local Suppression in Statistical Disclosure Control and Data Editing, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., and Willenborg, L.C.R.J. (1995b). Optimum Global Recoding and Local suppression in Microdata Sets. Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Nemhauser, G.L., and Wolsey, L.A. (1988). *Integer and Combinatorial Optimization,* Wiley, New York.

Pieters, A.J. and De Waal, A.G. (1995). A Demonstration of ARGUS, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Van Gelderen, R. (1995). ARGUS: Statistical Disclosure Control of Survey Data, Report, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

USING NOISE FOR DISCLOSURE LIMITATION OF ESTABLISHMENT TABULAR DATA

B. T. Evans and L. Zayatz[1]

ABSTRACT

The Bureau of the Census is looking into new methods of disclosure limitation for use with establishment tabular data. Currently we use a strategy that suppresses a cell in a table if the publication of that cell could potentially lead to the disclosure of an individual respondent's data. In our search for an alternative to cell suppression that would allow us to publish more data and to accommodate more requests for special tabulations, we are considering adding noise to our underlying microdata. By perturbing each respondent's data, we can provide protection to individual respondents without having to suppress cell totals. The question remains, however, as to the utility of the data after noise is added. In this paper we discuss the benefits and drawbacks of adding noise to microdata prior to tabulation, with respect to both disclosure issues and the behavior of published estimates.

KEY WORDS: Confidentiality; Disclosure; Noise; Cell Suppression.

1. INTRODUCTION

The responding unit in many economic surveys and censuses conducted by the Census Bureau is the establishment. Individual establishments' responses are weighted (where appropriate) and aggregated, and estimates are generally produced by categorical variables like Standard Industrial Classification (SIC) code or geography. Given the geographic information and other characteristics on which tables are based, in conjunction with general knowledge and publicly available sources, it is generally a reasonable assumption that the set of establishments contributing to a cell is well known to data users.

The Census Bureau collects information from respondents under Title 13, U.S. Code, which prohibits the Census Bureau from releasing "any publication whereby the data furnished by a particular establishment or individual under this title can be identified." The disclosure limitation problem is to prevent data users from being able to recover any respondent's reported values using values appearing in the published tables. The Census Bureau must ensure that a cell value does not closely approximate data for any one respondent in the cell and, moreover, that one respondent or a coalition of respondents cannot subtract their contribution(s) from the cell value to achieve a "close" estimate of the contribution of another respondent (Cox and Zayatz, 1993).

2. CELL SUPPRESSION

The Census Bureau's current disclosure limitation technique used for establishment tabular data is cell suppression. Cells that pose a disclosure risk are identified using one of two rules --- the *n-k* rule or the *p%* rule (see Federal Committee on Statistical Methodology, 1994 for a detailed explanation of these rules). All cells that fail the disclosure rule are called primary suppressions, or sensitive cells.

Cell suppression limits disclosure by removing from publication (suppressing) all sensitive cells plus sufficiently many additional cells, called complementary suppressions, to ensure that the values of the primary suppressions cannot be narrowly estimated through

[1] B. Timothy Evans and Laura Zayatz, Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA.

manipulation of additive relationships between cell values and totals (Cox and Zayatz, 1993). When a cell is suppressed, its total value is removed from the cell and replaced with a 'D' flag.

While the concepts behind determining whether a particular cell is a disclosure risk are relatively simple, the process of choosing complementary suppressions to protect these sensitive cells is very complicated. The methodology by which complementary suppressions are chosen, as well as the accompanying computer software, is very difficult to understand for anyone without a background in linear programming. Because of the structure of the computer programs, the process must be performed separately for each data product. Among other things, this means that analysts must keep track, from one data product to the next, of which cells have previously been suppressed (and hence must be suppressed and protected in all subsequent data products) and which cells have previously been published (and hence can't be used as complementary suppressions).

Coordinating suppression patterns among tables becomes practically impossible in the presence of multiple special tabulations. To truly prevent any disclosures, we would have to keep track of *all* special tabulations requested by *all* data users, identifying not only those cells that were suppressed in any of the tabulations but also any unsuppressed cells that could be used in conjunction with unsuppressed cells from another tabulation to recover the value of a suppressed cell. Thus we would have to keep track of all interrelationships between all cells across all publication tables and special tabulations. We simply do not have the resources to do this.

Another major drawback of cell suppression is that it suppresses much information that is not at risk for disclosure. Any cell that is used as a complementary suppression but that is not itself a primary suppression represents information that could have been published if there were some other way of protecting the sensitive cells. Particularly at fine levels of detail, the need for complementary suppressions often results in tables full of D's. Data users frequently complain that we suppress too much data.

3. INTRODUCTION OF NOISE TO MICRODATA PRIOR TO TABULATION

3.1 General description

An alternative way to protect individual respondents' data is to add noise to it. Suppose we perturb each respondent's data by a small amount, say 10%. Then if a cell contains only one establishment, or if a single establishment dominates the cell, the value in the cell will not be a close approximation to the dominant establishment's value because that value has had noise added to it. By adding noise, we would avoid disclosing the dominant establishment's true value.

If the noise is added in a careful, systematic way, we can also minimize its effect on cells that would not be suppressed under the usual procedure. Thus we can protect individual establishments without compromising the quality of our estimates.

To each establishment in our sampling universe we would assign a multiplier, or noise factor. Whenever an establishment was canvassed in any survey or census, all of its values would be multiplied by the establishment's assigned noise factor. Within a particular survey or census, all establishments would have their values multiplied by their corresponding noise factors before the data were tabulated. Note that because the same multiplier would be used with an establishment wherever that establishment was tabulated, values would be consistent from one table to another. That is, if the same cell appeared on more than one table, it would have the same value on all tables.

3.2 The multipliers

To perturb an establishment's data by about 10%, we would multiply its data by a number that was close to either 1.1 or 0.9. We could use any of several types of distributions from which to choose our multipliers. For instance, to perturb an establishment's data in a positive direction, we could choose the multipliers from a normal distribution with mean 1.1 and very small variance, perhaps .05 or .01. In any case, we would want to use a distribution with 1.1 as the measure of center (mean, mode, etc.) and with small variance about 1.1. If we wanted to ensure that all multipliers were at least 1.1, guaranteeing at least 10% protection to single-establishment cells, we could simply truncate our distribution at 1.1 and discard the portion below 1.1.

Whatever distribution we decide to use for generating multipliers near 1.1, it is of paramount importance that we use the same distribution, or rather its "mirror image," to generate multipliers near 0.9. In other words, if we consider the two distributions together, the overall distribution of the multipliers should be symmetric about 1. The reason for this condition is discussed in Section 3.3.

Under current practices, the unit of analysis for disclosure avoidance is the *company*. That is, we seek to protect respondent data at the company level as well as for individual establishments within the company. Because company-level values must be protected, all noise for a single company should either inflate or deflate that company's true values. In other words, all establishments from the same company should be perturbed in the same direction and hence have approximately (but not exactly) the same multiplier. This way, if all of the establishments contributing to a cell belonged to the same company, the resulting cell estimate would be perturbed by about 10%. Otherwise, the cell estimate could be very close to the company's true value if the noise in the positively-perturbed establishments (multipliers > 1) and the noise in the negatively-perturbed ones (multipliers < 1) happened to roughly cancel each other out. Thus by perturbing all of a company's establishments in the same direction, we ensure that company-level data is protected.

Note that different establishments belonging to the same company should not have *exactly* the same multiplier. In some cases, different establishments owned by the same company compete with each other. Suppose two establishments from the same company each appeared in cells by themselves. One of the establishments could use its own true value and the value in the cell in which it appeared alone to derive its own noise multiplier. Then, if the establishment assumed that the same multiplier was used with all establishments belonging to its parent company, it could derive the other establishment's true value by using the known multiplier and the value in the cell in which the other establishment appeared alone. Thus in order to protect establishments from each other within the same company, each establishment in the company should have a slightly different multiplier.

3.3 Assignment of multipliers and its effect on estimates

We would like to assign the multipliers in such a way that we minimize the effect of the noise on those cells that are not at risk for disclosure. In particular, estimates at higher levels of aggregation are not generally sensitive. In economic censuses and surveys, the most common sub-national estimates are produced by SIC, geography, or measure of size. In assigning the multipliers, we would like to arrange for these estimates to contain as little noise as possible.

One way to accomplish this is to ensure that among all establishments contributing to one of these estimates, the number having a positive amount of noise (multiplier > 1) and the number having a negative amount (multiplier < 1) are roughly equal. More precisely, considering that different establishments have different measures of size, we would like to ensure that the *absolute amount* of noise added and the *absolute amount* of noise subtracted roughly cancel each other out when summed over all establishments contributing to the estimate.

To this end, we would sort establishments by SIC × geography × measure of size before assigning multipliers. The first establishment would be assigned a multiplier close to 1.1; the second establishment would be assigned a multiplier close to 0.9; the third establishment 1.1; the fourth, 0.9; etc. This procedure is "better" than assigning noise randomly because it assures that for every establishment that is assigned a noise factor greater than 1, there is in general another establishment of about the same size in the same SIC and the same geographical area that is assigned a factor less than 1. Thus when aggregate estimates are computed, the noise present in these two establishments should have a tendency to cancel out.

Theoretically, both random assignment and systematic assignment of noise factors should provide that the expected value of the amount of noise present in any estimate is zero, thanks to the symmetry of the distribution of the multipliers. However, the systematic procedure takes advantage of the hidden stratification inherent in the universe of establishments and should help to reduce the *variance* of the amount of noise as compared to random assignment.

For other non-sensitive cells, we would still have the result that, on average, estimates would not be altered by much. For aggregate estimates not computed along SIC × geography × measure of size lines, for detailed cells with many contributors, and for detailed cells having few contributors but of roughly the same size, it is still true that the establishments that are perturbed in the positive direction and those that are perturbed in the negative direction will generally balance each other out. Most of these cells should end up with little noise, although we can't ensure this as effectively as for the other aggregate estimates. For these estimates, the systematic assignment of noise factors would be similar to random assignment in terms of the average amount of noise present.

In contrast, a cell that is dominated by a single contributor would most likely contain a large amount of noise. If the largest contributor is very large compared to all others in the cell, it is much less likely that positively-perturbed establishments and negatively-perturbed establishments will cancel each other out when determining the amount of noise present in the cell estimate. Looked at another way, the more dominant the largest contributor, the more the amount of noise present in the cell estimate will resemble the amount of noise present in the largest contributor (about 10%). Thus the cells that would have been at greatest risk for disclosure would in general receive the most noise.

3.4 Adding noise to survey data

In surveys, each respondent's data is generally weighted inversely proportionally to the establishment's probability of being included in the survey sample. For establishments with large weights, the weight itself offers some protection against disclosing the respondent's actual reported values. For survey data, to reflect the protection already provided by the sample weight, noise would be applied as follows:

For each establishment in a cell, calculate

$$\text{establishment value} \times \left[\text{multiplier} + (\text{weight} - 1)\right]$$

and then add up these noise-added establishment values to obtain total cell value. Note that noise is added only to one multiple of each establishment's value, and the remaining (weight - 1) multiples have no noise added.

This procedure has the effect of changing (weighted) values for certainty or near-certainty establishments (those having weights close to or equal to 1) by a large amount while changing weighted values for establishments with large weights by a small amount. This is desirable because we are more concerned with the disclosure risk of certainty establishments, whose values aren't protected by their weights.

3.5 Updating multipliers

Many surveys produce trend statistics, statistics that indicate a percent change in the level of a variable from one time period to another. If we were to use the same multiplier for the same establishment in successive iterations of a periodic survey, we would be showing exact percent changes for establishments in single-establishment cells, and this would be a disclosure. (The common multiplier factors out of all level estimates, yielding the true percent change.) Multipliers would have to be changed from one period to the next in order to protect these trend statistics.

In order to preserve the utility of trend statistics, we would update multipliers in such a way that a particular establishment's values were always perturbed in the same direction in successive iterations of the survey. In other words, if the original (first) multiplier was chosen from a distribution centered at or near 0.9, we would choose all new multipliers from the same distribution. If the original multiplier was close to 1.1, we would choose all new multipliers close to 1.1. This way, if a particular estimate ends up being biased because of the addition of noise, in spite of our efforts to the contrary (see Section 3.3), it will at least be biased by roughly the same amount from one period to the next, thereby preserving the trend. Otherwise, if the direction of the bias were able to change from one period to another, the underlying trend might be obscured by the noise.

By using similar multipliers with a single establishment from one period to the next, we would maintain the longitudinal qualities of the data. And by varying the noise factors slightly between periods, we would avoid disclosing the exact value of any establishment's true percent change.

3.6 Using different multipliers for different data items

In addition to protecting the values of an establishment's individual data items, we also need to protect the relationships between data items. For example, if in some survey an establishment reports its total revenue as well as components of the total like advertising revenue, we would need to protect the ratio of advertising revenue to total revenue for that establishment. This would only be a concern in single-establishment cells; as long as there were two or more establishments contributing to a cell, it would be impossible to distinguish any one establishment's exact share of each data item. In a single-establishment cell, however, it is known that 100% of each data item (with or without noise) is attributable to the sole contributor. Thus if the data items in both the numerator and the denominator of the ratio were multiplied by the same noise factor, then in computing the ratio the noise factor would cancel out of both the numerator and denominator, yielding the true ratio for the establishment.

To protect inter-variable relationships, we could use different multipliers for different data items. A base multiplier would be maintained for each establishment, and a different adjustment factor would be assigned for each item published. For example, say establishment A's base multiplier is 1.12 and establishment B's is 0.87. When tabulating total revenue, we might multiply establishment A's value by 1.123 and establishment B's value by 0.867. An adjustment of 0.003 has thus been added to (or subtracted from) the base multipliers for purposes of tabulating total revenue. When tabulating advertising revenue, we might multiply establishment A's value by 1.125 and establishment B's value by 0.865, in which case an adjustment of 0.005 has been added to (or subtracted from) the base multipliers. Thus when computing the ratio of advertising revenue to total revenue for either establishment, the different adjustments to the base multiplier for different data items would prevent exact disclosure of the ratio.

A major drawback to using different multipliers for different data items is that we could no longer guarantee that detail data items added to their proper totals within an establishment. One possible solution would be to define one selected detail data item as the difference between the aggregate data item and the sum of all other detail items. This would guarantee additivity but would have an unpredictable effect on the data item selected to be defined as the difference.

3.7 Flagging cells with a large amount of noise

All resulting table cells containing a large percentage of noise, say a 7% change in value or more, would be flagged so users would know that the values may not be useful. This set of cells would encompass most sensitive cells, as well as a few non-sensitive cells that received a lot of noise simply through randomness. The description of the flag would explain how and why noise was added and would let users know that disclosure limitation had been performed. We could also use the same flag on any cells that were identified as sensitive but that, because of randomness of multipliers, did not receive much noise. In this case, the users would at least *think* the cell contained a lot of noise and would hesitate to treat the cell value as reliable. We would expect relatively few cells of this type.

Cells having too much noise, as well as sensitive cells not sufficiently protected by the noise, would contain a flag but no published value. The value of the cell may still be derivable, but the fact that the value does not actually appear in the cell would draw attention to the fact that we don't consider the estimate reliable. This is similar to how we treat cells having high coefficients of variation (CVs) in survey publications. By not publishing actual values, we may also lessen the *appearance* of disclosure for single-establishment cells and for sensitive cells that did not receive much noise.

4. BENEFITS OF NOISE

4.1 Simple procedure

Adding noise to establishment-level data before producing tables has several advantages over the traditional cell suppression techniques. First, it is a far simpler and less time-consuming procedure than cell suppression. Each establishment would need only to have its data items multiplied by the establishment's noise factor, possibly with different adjustments to the noise factor for different data items, prior to tabulation. In each table, we would still have to identify those insufficiently-protected sensitive cells (cells that would normally be primary suppressions) in order to flag them, but the complicated and lengthy process of choosing complementary suppressions would be avoided. The addition of noise would not be table-specific, whereas complementary suppressions must be identified on a table-by-table basis. Computer programs for adding noise would also be much easier to write, modify, run, and understand than the programs that currently exist for choosing cell suppression patterns.

4.2 Simplifies release of multiple data products

Another important advantage of adding noise is that it would eliminate the need to coordinate cell suppressions between tables. Under the current cell suppression practices, disclosure analysis must be done separately for each data product. This involves keeping track, from one data product to another, of all cells that have previously been published and all cells that have previously been suppressed. (Otherwise, for instance, a cell that is used as a complementary suppression on one table might appear unsuppressed on another table.) Keeping track of suppressions is difficult to orchestrate and difficult to understand. However, using noise to protect estimates would make this unnecessary. For each data release, we would need only to identify and flag any cells that were primary suppressions or that contained more than the prescribed acceptable level of noise.

In particular, eliminating the need to coordinate suppressions would allow for easy and quick fulfilment of requests for special tabulations. Using noise to protect estimates would allow us to compute as many special tabulations as we needed without having to keep track of what estimates had already been released. We would again need only to identify and flag any primary suppressions and cells containing too much noise.

4.3 Allows for publication of more data in standard releases

The addition of noise was designed mainly to overcome the multiple special tabulation problems that arise with cell suppression, but it also may allow for more valuable data to be published in our standard releases. With cell suppression, users lose information both for cells which are primary suppressions and for those that are complementary suppressions. With the noise technique, the sensitive cells (normally primary suppressions) would in general receive a lot of noise and be flagged as such. In contrast, non-sensitive cells would receive little noise, including some that would have been complementary suppressions. Thus for publications which normally contain many complementary suppressions, the noise technique should provide data users with more valuable information.

Note that adding noise would not help much with tables where most suppressions are primary suppressions. For those tables, only a reduction in detail could reduce the number of cells for which data was suppressed or severely altered.

5. ARGUMENTS AGAINST NOISE

5.1 Insufficient protection for single-establishment cells

Some respondents may not feel that the added noise provides enough protection to values in single-establishment cells. If a cell in a table has only a single establishment contributing to it, cell suppression would suppress the cell's value and the cell would simply contain a 'D'. Under the noise approach, the cell would contain a flag noting that the value in the cell had been severely altered, but the actual value may still be derivable through subtraction. The flag may lessen the *appearance* of disclosure, since no value would appear in the cell. However, the respondent may still feel uneasy about the derived number being an estimate of his actual value, even if the estimate contains a lot of noise and is flagged as being unreliable. The suppression approach may give the appearance of offering more protection.

On the other hand, for every value that is suppressed under the cell suppression approach, an interval which contains that value can be derived. For example, if a value of 100 is suppressed, users can look at surrounding cells to determine that the value is between, say, 84 and 124. Users often derive this interval and then use the midpoint as an estimate for the cell value, in this case 104. Sometimes the midpoint is very close to the true value, and other times it is not.

Which method offers the best protection? This is a subjective question and is not easily answered. The point is that the question of whether noise provides *enough* protection to sensitive cells could just as easily be directed at the cell suppression approach.

5.2 Perceptions of data quality

It is possible that some respondents may resent putting time into preparing good responses if they know the Census Bureau is only going to add noise to them anyway. We would need to emphasize that we were not simply adding noise indiscriminately. Noise would be added in an unbiased, controlled way so as to preserve the statistical properties of the data while having a negligible effect on non-sensitive estimates. To maintain the properties of the data, we would need to know what those properties were to begin with. And to assess the effect of the noise on important aggregate estimates, we would have to know their true noise-free values as accurately as possible. Thus it is crucial that we begin the noise addition process with the true values in order to perturb the data in a predictable way.

There also may be concern on the part of some data users as to the quality of the data after noise has been introduced. In using this proposed technique, we would hope that the users' desire for multiple special tabulations and their desire to see more published cells (at the expense of noise) would outweigh their desire for true values (at the expense of suppressions). Flags would inform users of the data's utility by drawing their attention to cells that had been adversely affected by the noise. Also, users know that our published estimates for surveys already have sampling error associated with them, as described by the CV, and are not true, exact values. Even our census values contain some "noise" due to various types of nonsampling errors (reporting errors, keying errors, imputation, etc.). In general, users know that we are publishing our best possible estimate of each cell's value.

The fact that we would be deliberately perturbing the data, however, may lead users to feel that the numbers they see in the tables are not our best possible estimates. While the other types of error that are already present in our published values are errors that we attempt to control or eliminate, the added noise would be error that we were *actively introducing* into our estimates. We would have to emphasize the purpose of adding noise and remind users that it was added in a way that would minimize its effect on non-sensitive estimates.

6. VARIATIONS ON ADDING NOISE

6.1 Noise with some cell suppression

As mentioned in Section 5.1, many people feel uncomfortable about the idea of publishing any value, even one with a lot of noise in it, for cells with only 1 or 2 establishments. To address this issue, they have asked if noise could be applied *and* those particular cells could still be suppressed, along with a sufficient number of complementary suppressions. (Other cells that were identified as sensitive but that had 3 or more establishments in them would not be suppressed and would be protected by the noise and the accompanying flag.)

We could suppress one- and two-establishment cells, but there are several disadvantages to this approach. Two procedures would have to be applied to the data, thus making disclosure limitation more time-consuming and harder to understand. There would still need to be suppression pattern coordination among all tables and special tabulations, although there would be fewer suppressions to coordinate. This approach seems to possess the disadvantages of both cell suppression and noise, and it does not solve our problem with multiple special tabulations.

The fact that only a flag, and *not* a value, would appear in the cell may help reinforce the idea that the cell's value is protected by the noise. The decision as to whether the noise and the flag offer *enough* protection, however, may ultimately rest with the respondents.

6.2 Cell suppression for standard releases and noise for special tabulations

Many people have asked about the possibility of using cell suppression for all standard publications or tables and using noise for all special tabulations. This practice could compromise the protection provided by cell suppression. A special tabulation could contain some of the same cells that appeared in a standard publication. For cells that were primary suppressions in the standard release, this would not be a problem because in the special tabulation these cells would contain a lot of noise and would be flagged. However, cells that were suppressed purely as complementary suppressions in the standard release should not receive much noise in the special tabulation and thus would not be flagged. A user could substitute the values from the special tabulation, which would be relatively noise-free, into the corresponding suppressed cells in the standard table and through addition and subtraction could obtain close approximations of some primary suppressions. Thus we would lose much of our protection for the primary suppressions.

Another problem with using noise only in special tabulations is inconsistency between tables. If a special tabulation contained some of the same cells that appeared in a standard publication, these cells would contain noise in the special tabulation but not in the publication. It would be inconsistent to have the same cell appearing in two locations with a different value in each location.

6.3 Adding noise and raking to true values

6.3.1 General strategy

One of the main objections to the idea of adding noise is that it would affect *all* estimates, not just those that would have been disclosure risks. An alternative that would addresses this problem is to add noise but then force the published values of the more important estimates (presumably those at higher levels of aggregation) to equal their true (without noise) values. Interior cells would then be raked, or proportionally adjusted, so that they still summed to the aggregate estimates.

We would first need to determine the level of aggregation at and above which we wanted published estimates to equal their true values, i.e., their values uncorrupted by noise. This should be a level at which we expect very few, if any, sensitive cells. We would proceed to introduce noise into all establishments and compute all estimates with noise present. Then, at the lowest level of aggregation at which published estimates were to equal their true values, we would force the estimates with noise present to equal the corresponding values with no noise, raking the interior cells as well to preserve additivity and proportions. For multi-dimensional tables, the interior cells would have to be simultaneously raked to all marginal totals that were held fixed at their true values. All estimates at levels of aggregation higher than the level at which we raked would automatically equal their true values, since they would be summations of components that had already been raked to their true values.

This method would leave many non-sensitive cells, namely those in which true values would be published, totally unaffected by noise. At the same time, the raking should not have a very pronounced effect on any of the other estimates. Since estimates at higher levels of aggregation, where the raking would be done, shouldn't have much noise in them to begin with, the raking factors would be relatively small. Those cells that had a lot of noise before raking would still have a lot of noise after raking, so the noise would still offer protection to sensitive cells.

There are two ways to approach the raking, as described in the next two sections. One option is to rake each table individually; the other is to rake to all fixed marginal totals simultaneously before producing any tables. Each has advantages and disadvantages as compared to the other.

6.3.2 Raking each table separately

One approach to raking is to rake each data product individually. This has the advantage of allowing analysts to determine on a table-by-table basis (including special tabulations) the lowest level of aggregation at which the estimates will not be disclosure risks. The interior cells in a particular table would then be raked only to the selected marginal totals that actually appeared in the table, regardless of what other marginal totals might be held fixed in other tables. This would in general allow more cells to equal their true values than in the case where estimates are simultaneously raked to all marginal totals that are to be held fixed. This is described in Section 6.3.3, and is related to the level of detail in the table.

An establishment would retain the same *base* noise factor throughout all tables. However, since each table would be raked individually, the *net* amount of noise (after raking) present in an establishment's contributions to different cells could differ from one table to another, and even from one column (or row) to another within the same table. This could create problems of consistency between tables. One requirement to maintain consistency is that if a group of cells are published with no noise added in one table, they should be noise-free in all tables in which they appear. Otherwise the same cell could have different values in different tables. We would therefore have to keep track, from one table to the next, of what cells had previously been held fixed. This includes keeping track of noise-free cells when producing special tabulations. This process would be similar in complexity to keeping track of suppression patterns among tables under the cell suppression scheme. If we were to use this technique, it would probably be best reserved for small-scale surveys having tables that were relatively non-interrelated.

6.3.3 Raking once before tabulating

One way to avoid having the same cell appear in two different tables with two different values is to rake all estimates only once. Before producing any tables, analysts would determine the set of cells that would be forced to equal their true values. Then all other cells from all publication tables would be simultaneously raked to all of these fixed cells.

This single raking would be done by first constructing an $n \times n$ "supermatrix," where n is the total number of categorical variables that appear in any of the tables. After adding noise to all establishments, each establishment would then be tabbed into *exactly* one interior cell of this supermatrix, depending on its values of the n categorical variables. Then, an n_0-dimensional raking would be done, where n_0 is the number of categorical variables for which some or all of the marginal totals were to be fixed at their true values ($n_0 \leq n$). The raking would define an adjustment factor for each cell, indicating the percentage by which the raking changed the value in the cell. This raking factor would be applied to each establishment contributing to the cell. The net noise factor for the establishment would then be the product of the original noise factor and the adjustment factor determined by the raking.

This net noise factor would then be associated with the establishment throughout the production of all standard tables and special tabulations. This would guarantee consistency of estimates between tables because the same set of establishments in a cell with the same set of noise factors would always produce the same estimate. And producing special tabulations would require no special procedures; we would simply tabulate each establishment's value, multiplied by the single factor that reflects both noise and raking, and then flag appropriate cells. We would not have to worry about keeping track of what estimates had appeared in previous tables and how much noise they contained.

The main disadvantage of the single-rake approach is that it would limit the level of detail at which we could force estimates to equal their true values. In the presence of a large number n of categorical variables, interior cells in the supermatrix would generally be sparsely populated. If we tried to rake the interior cells to too many fixed marginal totals simultaneously, some cells may then be constrained to equal their true values to guarantee (sometimes trivial) additivity in all dimensions of the supermatrix, thereby cancelling the effect of the noise. Analysts would have to limit the number of marginal totals that were held fixed; otherwise, the raking could undo whatever protection was provided to interior cells by the addition of noise, which could leave many sensitive cells unprotected.

7. CONCLUSIONS

Adding noise to establishment microdata has clear advantages over cell suppression as a way of providing the required protection to individual respondents. As we move into an era of customized data products and user-defined tables, the noise technique would afford us the flexibility to accommodate a wide variety of data requests without the worry of inadvertently disclosing any particular respondent's values. Unlike with cell suppression, we wouldn't have to keep track of all prior requests in order to guarantee that each new data product was free of disclosures.

The strongest argument against the noise approach is that it would affect all estimates, not just those requiring protection. Several modifications to the approach have been suggested to address this issue. Of these, the idea of adding noise and then raking to true values at higher levels of aggregation seems to hold the most promise; it is this option that we intend to investigate further. If raking proves to be a satisfactory response to the question of data quality in the presence of noise, the addition of noise to microdata could quite possibly become the Census Bureau's preferred technique for disclosure avoidance with establishment tabular data.

Note that a more detailed version of this paper will appear in the Census Bureau's Statistical Research Division Report Series (Evans and Zayatz).

REFERENCES

Cox, L.H., and Zayatz, L. (1993). Setting an agenda for research in the Federal Statistical System: Needs for statistical disclosure limitation procedures, *Proceedings of the Section on Government Statistics, American Statistical Association*, 121-126.

Evans, B.T., and Zayatz, L. (1996). Using noise for Disclosure Limitation of Establishment Tabular Data, *Statistical Research Division Report Series*, Bureau of the Census, to appear in 1996.

Federal Committee on Statistical Methodology (1994). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC: U.S. Office of Management and Budget.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

ASSESSMENT AND REDUCTION OF DISCLOSURE RISK IN MICRODATA FILES CONTAINING DISCRETE DATA

J.-R. Boudreau[1]

ABSTRACT

Estimating the number of unique elements in a population from a sample, and in particular estimating the conditional probability of being a unique element in a population given being a unique element in the sample, is essential to the development of confidentiality rules for all types of products for dissemination. For the case of microdata based on discrete variables, we determine the exact relationship between the unique elements in the population and those in the sample. In addition, we give an unbiased estimator of the number of unique elements in the population. Its great sampling variability for small sampling fractions forces us to consider modelling this relationship. After observing this conditional probability for a number of real populations, we provide a parametric formulation of it. This formulation is only empirical; it has no theoretical justification. However, this model can be used to develop confidentiality rules. We conclude this article by describing a method of identifying records that are potentially problematic. This technique enables designers of microdata to introduce noise into data only in records where this is really necessary.

KEY WORDS: Disclosure risk; Confidentiality; Microdata; Uniqueness; Identification.

1. INTRODUCTION

Consider the following matching problem. A simple random sample (File A) is drawn from a population. We want to match this file with another file (File B) from the same population, using all the discrete variables common to the two files. We assume that capture and response errors are negligible. If a one-to-one match is obtained between two records, what is the confidence level that we can attribute to the statement, "These two records are from the same unit in the population"?

The author sees an immediate application here. The above confidence level can be used to assess the disclosure risk of File A. Say that a statistics agency disseminates a microdata file A. An individual or company, possessing a microdata file B that has both a unique key (e.g., names and addresses) and a few variables in common with A, can carry out a match with the latter in order to identify the origin of certain records. In light of this, the statistics agency must make sure before disseminating its file that the confidence level is as low as possible, so as to remove any incentive to match the disseminated file with other files.

A necessary condition for having a high confidence level is that File B must cover the entire population. According to this hypothesis, the confidence level is directly related to the conditional probability of being a unique element in the population (in relation to the matching variables) given being a unique element in the sample. As we will see below, determining this probability is in part related to estimating the number of unique elements in the population on the basis of the sample.

In recent years, much research has been done on estimating the number of unique elements in the population. Greenberg and Zayatz offer two ways to estimate the number of unique elements. The first way is to redo the sample according to the same sample design. The estimator is constructed assuming that the relationships between the unique elements in the population and the first sample are the same as between those of the first sample and the second. The second way proposed by these authors uses the population structure, that is, the description of the population in terms of the number of cells defined by the matching variables having exactly one unit, two units, etc., which they call

[1] Jean-René Boudreau, Statistics Canada, Social Survey Methods Division, 15-P, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

"classes of equivalences." We will use this technique as a point of departure. These two techniques yield good results if the sampling fraction is greater than 10%. Another way to proceed is to try to model the population structure on a sample. Bethlehem *et al.*, tried to model the proportion consisting of the number of unique elements in the population using the model derived from the Poisson-Gamma distribution. This model suffers from a major lack of adjustment. Skinner modelled the proportion of unique elements in the population by using the Poisson-Lognormal distribution. The results that he obtained correspond much more closely to reality.

The author proposes to use sampling theory to determine exactly the form of the relationship between the unique elements in the population and those in the sample. We will show that it is practically futile to try to solve the problem for small sampling fractions by using only sampling theory. Consequently we will try to model this relationship for small sampling fractions. We will also offer a method, admittedly still at the conjecture stage, for identifying upper limits for the conditional probability, which will thus enable us to control the disclosure risk for a microdata file.

2. DETERMINING CONDITIONAL PROBABILITY

Say that we have a population of N elements or units. The content, that is, the matching variables, partitions this population into m subpopulations of size $N_1, ..., N_m$. The structure of the population is given by the vector $(U_1, ..., U_N)$ where U_j = card $\{k\colon N_k{=}j\}$. We take a simple random sample of size n from this population. We observe the random vector $(n_1, ..., n_m)$, the components of which are respectively the number of units sampled in the subpopulation k $(k = 1, ... ,m)$. The structure of the sample is the random vector $(u_1, ..., u_n)$ where u_j = card $\{k\colon n_k{=}j\}$.

An element will be said to be unique in the population if it belongs to a subpopulation the size of which is unity. A sampled unit will be said to be unique in the sample if it is the only sampled unit to belong to its subpopulation. Since if a unique element in the population is sampled it is necessarily unique in the sample, we obtain that the conditional probability of being unique in the population given being unique in the sample is the ratio between the proportions of unique elements in the population and in the sample. Thus we want to have an estimate of

$$P = \frac{\dfrac{U_1}{N}}{\dfrac{E\{u_1\}}{n}} = f\,\frac{U_1}{E\{u_1\}}$$

where f is the sampling fraction and the mathematical expectation is the established by the sample design. The expectation is necessary in order to obtain a parameter at the level of the population. This parameter, which (although this is actually a misnomer) will be considered as a conditional probability, is not far from the idea of disclosure risk or the confidence level explained in the previous section. We have a first result.

Theorem A. *If a simple random sample of size n is drawn from a population of size N that has the structure* $(U_1, ... , U_N)$, *then*

$$E\{u_j\} = \frac{\binom{N-j}{n-j}}{\binom{N}{n}}\,U_j + \sum_{i=1}^{\infty} \frac{\binom{j+i}{j}\binom{N-j-i}{n-j}}{\binom{N}{n}}\,U_{j+i}\,.$$

Demonstration. The sum is in reality finite. Since the values of u_j are integers, we can use the identity

$$E\{u_j\} = \sum_{i=1}^{\infty} P\{u_j \geq i\}\ .$$

Say that $A_k = \{(n_1, ..., n_n)\colon n_k{=}j\}$. We have the following identity:

$$P\{u_j \geq i\} = P\left\{\bigcup_{k_1 < \cdots < k_i} A_{k_1} \cdots A_{k_i}\right\}.$$

It can easily be shown that

$$\sum_{i=1}^{\infty} P\{u_j \geq i\} = \sum_{k=1}^{m} P\{A_k\}\ .$$

All that we need to do is determine the probability of each union and realize that all the terms cancel each other out except for the sum of the probabilities of events A_k. Now, $P\{A_k\}$ is equal to

$$P\{A_k\} = \frac{\binom{N_k}{j}\binom{N-N_k}{n-j}}{\binom{N}{n}}\ .$$

Thus the expectation of u_j is equal to

$$E\{u_j\} = \sum_{\substack{k=1 \\ j \le N_k \le N-n+j}}^{m} \frac{\binom{N_k}{j}\binom{N-N_k}{n-j}}{\binom{N}{n}}$$

$$= \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N-i}{n-j}}{\binom{N}{n}} U_i \ .$$

Which it was necessary to demonstrate.

In particular, for $j = 1$, we have

$$E\{u_1\} = f\, U_1 + \sum_{i=1}^{\infty} (i+1) \frac{\binom{N-1-i}{n-1}}{\binom{N}{n}} U_{1+i} ,$$

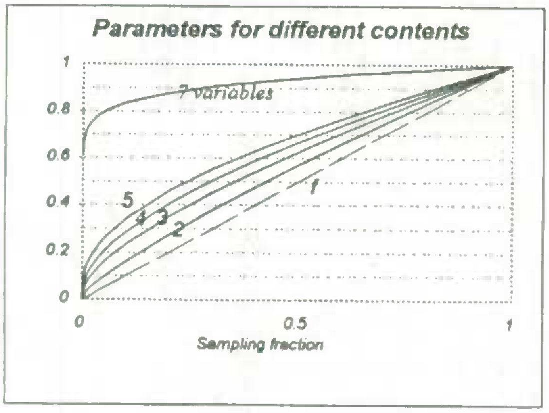
which may be written as

$$E\{u_1\} = f\, U_1 + n \sum_{i=1}^{N-n} \frac{(i+1)}{N-i}\left(1 - \frac{n}{N}\right) \cdots \left(1 - \frac{n}{N-i+1}\right) U_{1+i}$$

$$\approx f\left(U_1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^i\, U_{1+i}\right)$$

if N is sufficiently large. Thus the conditional probability becomes

$$P = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^i \,\frac{U_{1+i}}{U_1}} \ .$$



This graph gives the relationship between P and the sampling fraction for different contents. The size of the population is nearly 800,000. We have retrieved seven variables from this population. The first content is defined by the first two variables retrieved, the second is defined by the first three variables, and so forth. The greater the number of variables, the higher the conditional probability. The area of interest is undeniably the interval [0, 0.1]. The value of P at the origin is given by setting $n = 1$ in the exact formula or by setting $f = 0$ in the approximate formula of P. The value of P at this point gives $P = U_1/N$, which is exactly the proportion of unique elements in the population. The behaviour of the curve at the origin is analysed by developing P around 0. To do so, we will introduce the probability distribution $p(\cdot) = (p_1, p_2, \ldots)$ defined by

$$p_i = \frac{i\ U_i}{N}$$

for $i \geq 1$. If we denote the expectation of this distribution by $E_p\ \{\cdot\}$, the conditional probability may be written in the form

$$P = p_1 \frac{1 - f}{E_p\{(1-f)^X\}}$$

for all $f \leq \min_{1 \leq k \leq m} (1 - N_k / N)$ and $X(i) = i$.

In this form, it can easily be seen that the derivatives of a given order of P evaluated at the origin are sums of moments centred on the origin of the p distribution (with the denominator being the function that generates the moments). The first terms of the development are

$$P = p_1 + p_1(\mu_p - 1)\, f + \frac{p_1}{2}\left(\mu_p^2 - \mu_p - \sigma_p^2\right) f^2 + o(f^3) \ .$$

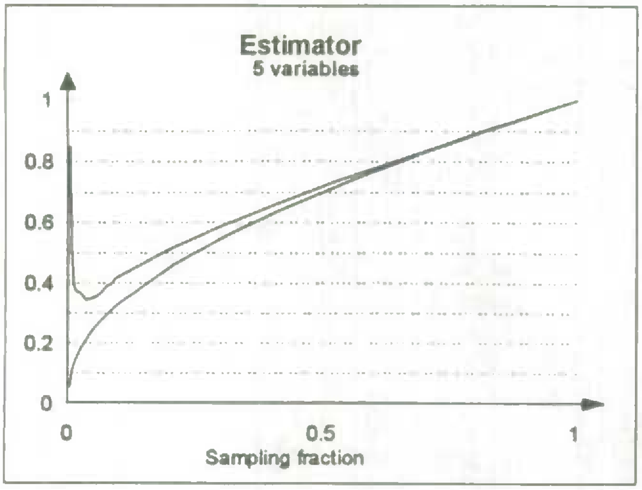
This expression tells us that P is increasing in the vicinity of 0 and is convex or concave, depending on whether the coefficient of variation of $p(\cdot)$ is smaller or larger than $1 - 1 / \mu_p$. It appears that a concave curve is a general fact for real populations. This statement is quite important. The relationship between P and f can result in just about anything if only the condition $N = U_1 + 2U_2 + \ldots + N\,U_N$ is verified. In fact, the latter condition is not sufficient in order for the population that has this structure to be able to be characterized as "real" or "observed". We do not know how the real populations are simulated, but it appears that the distribution of $p(\cdot)$ is responsible for the concavity of the relationship.

3. NATURAL ESTIMATOR OF *P*

The form of the expression for P suggests, as an estimator of this quantity, that the population structure should be replaced by the sample structure, meaning that the estimator would be

$$\hat{P} = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^i\,\frac{u_{1+i}}{u_1}},$$

an estimator consistent with the thinking of Greenberg and Zayatz.



The estimator converges toward 1 when the sampling fraction tends toward 0 instead of converging toward the proportion of unique elements in the population. This disastrous behaviour is due to the fact that the proportion of unique elements in the sample becomes disproportionate, owing to the smallness of the sample. In the section that follows, we will provide a measure of the divergence between the population structure and the sample structure for small sampling fractions. These results will force us to model P over this range.

4. UNBIASED ESTIMATOR OF U_j

To see how the sample structure grows more different from the population structure when the sampling fraction tends toward 0, we propose to construct an unbiased estimator for the components of the population structure. This result is found in Goodman. The divergence of the structures will be converted into the sampling variation of the estimator. To construct this estimator, we need the following lemma.

Lemma B. *For $1 \leq k \leq r$, we have the following identity*

$$\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \binom{r+i-1}{k-1} = \binom{r-1}{k-1}. \qquad (1)$$

Demonstration. We know that the left term of (1) is equal to

$$\frac{1}{(k-1)!} \sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \frac{i(i+r-1)\cdots(i+r-k+1)}{i} \qquad (2)$$

Say that $g(x) = x(x+r-1)\ldots(x+r-k+1)$. $g(\cdot)$ is a polynomial of degree k. We can thus apply the identity (Mercier, corollary 1, p. 141).

$$\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \frac{g(i)}{i} = g'(0) + g(0) \sum_{i=1}^{k} \frac{1}{i}.$$

Since $g(0) = 0$ and $g'(0) = (r-1)!\,/\,(r-k)!$, we obtain that (2) is equal to the binomial coefficient of the right term of (1). This completes the demonstration.

We have the following theorem.

Theorem C. *A simple random sample of size n is drawn from a population with the structure $(U_1, \ldots, U_N)$. If $n \geq N_k$ $(k = 1, \ldots, m)$, then an unbiased estimator of U_j is given by*

$$\hat{U}_j = \frac{\binom{N}{n}}{\binom{N-j}{n-j}}\,u_j - \frac{\binom{N}{n}}{\binom{N-j}{n-j}} \sum_{i=1}^{\infty} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}}\,u_{j+i}$$

Demonstration. According to theorem A, we have

$$E\{u_j\} = \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N-i}{n-j}}{\binom{N}{n}}\,U_i\,.$$

If we replace the u_j values by their expectation, we obtain

$$E\{\hat{U}_j\} = U_j + \sum_{k=1}^{\infty} \frac{\binom{k+j}{j}\binom{N-k-j}{n-j}}{\binom{N-j}{n-j}}\,U_{j+k}$$

$$-\sum_{i=1}^{\infty} \sum_{r=i+j}^{\infty} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}} \frac{\binom{r}{i+j}\binom{N-r}{n-i-j}}{\binom{N-j}{n-j}}\,U_r$$

$$= U_j + \sum_{k=1}^{\infty} \frac{\binom{k+j}{j}\binom{N-k-j}{n-j}}{\binom{N-j}{n-j}} U_{j+k}$$

$$- \sum_{k=1}^{\infty} \sum_{i=1}^{k} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}} \frac{\binom{k+j}{i+j}\binom{N-k-j}{n-i-j}}{\binom{N-j}{n-j}} U_{j+k} .$$

The binomial terms of the second sum simplify to give us

$$\sum_{k=1}^{\infty} \frac{\binom{k+k}{j}}{\binom{N-j}{n-j}} \frac{(N-j-k)!\,(k-1)!}{(N-n-1)!\,(n-j)!}$$

$$\left\{ \sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \binom{N-n+i-1}{k-1} \right\} U_{j+k} .$$

Which, according to lemma B, gives us

$$\sum_{k=1}^{\infty} \binom{j+k}{j} \binom{N-n-1}{k-1} \frac{(N-j-k)!\ (k-1)!}{(N-n-1)!(n-j)!} U_{j+k}$$

$$= \sum_{k=1}^{\infty} \frac{\binom{j+k}{j}\binom{N-j-k}{n-j}}{\binom{N-j}{n-j}} U_{j+k} .$$

The two sums cancel each other out, and we obtain the desired result.

If N is sufficiently large, we can use the following approximation

$$\hat{U}_j \approx f^{-j} \left\{ u_j - \sum_{i=1}^{\infty} (-1)^{i+1} \binom{j+i}{i} (f^{-1}-1)^i u_{j+i} \right\}$$

in place of the estimator. In particular, for $j = 1$, we obtain

$$\hat{U}_1 \approx f^{-1} \left\{ u_1 - \sum_{i=1}^{\infty} (-1)^{i+1} (i+1) (f^{-1}-1)^i \, u_{1+i} \right\} .$$

For $f < 0{,}5$, we clearly see that this estimator is unusable, since its sampling variance rises exponentially when N increases. Note that the condition $n \geq N_k$ for all k is necessary in order to have an unbiased estimator. If we keep this estimator even though the condition is not verified, a bias will be introduced that will not substantially reduce the root mean square error, since the problem lies with the size of the sample population. For example, for a population of 1,000,000 and a sampling fraction of 0.001, the sample size is 1,000. There is therefore a very strong possibility that the largest of the indexes i for which $u_{1+i} > 0$ will be such that the estimate will no longer mean anyththing (if $u_5 = 1$, one of the terms of the estimator will be of the order 10^{15}).

This shows that in order to estimate the conditional probability or a component of the population structure for small sampling fractions, the sample structure has little value. This throws cold water on those desiring a non-parametric solution. Let us now turn our attention to the possibilities of parameterizing the relationship between P and f.

5. PARAMETERIZATION OF THE CONDITIONAL PROBABILITY

In recent years, a number of persons have tried, with varying degrees of success, to model the population structure (in particular the number of unique elements in the population). The first attempt known to this author is that of Bethlehem *et al.*, . The latter assumed, without any justification other than the simplicity of the techniques, that the structure of a population could be simulated using a Poisson-Gamma model. According to this hypothesis, it is easy, using such a model, to find a parametric expression for the proportion of unique elements in the population. The expression is given by

$$E_m \{ U_1/N \} = \left(\frac{1}{1 + \beta N} \right)^{1+\alpha} ,$$

where α and β are parameters of the Gamma distribution of the model (α, $\beta > 0$). When we try, on the basis of a sample, to estimate these parameters, it quickly becomes apparent that the model suffers from a lack of adjustment. The parameter α is invariably estimated to have a value not significantly different from zero. Even the classical compensation techniques do not serve to stabilize the model. As will be seen below, the problem lies with the scope of the definition of α. Skinner *et al.*, proposes an approach based on classification theory. According to this theory, the population structure would be simulated by a Poisson-Lognormal model. This model is much more difficult to master than the one described above; in particular, the estimator of the proportion of unique elements in the population is the implicit solution of an integral equation that has no primitive. According to the results obtained by Skinner on Italian populations, this model appears to correspond closely to reality. The approach developed in this article seeks to directly model the relationship between P and f instead of modelling the underlying structure. This approach, which is a purely empirical one, has the advantage of corresponding to reality if one is able to observe a great number of real populations. However, a disadvantage of this method is that it does not give any information on how these populations are generated. In other words, it neither provides nor even suggests any theoretical justification.

This empirical approach does not rely on any probabilistic hypothesis except for that of selection of a simple random sample. The technique consists in studying the relationship between P and f for a number of populations obtained through censuses, trying to clarify the resemblances, proposing a parametric formulation of P as a function of f, and proposing a method of parameter estimation using a sample. The ultimate aim of this exercise is to propose a series of confidentiality rules that ensure that if the model holds up and if the estimation method is satisfactory, the lowest possible conditional probability is achieved.

5.1 Formulation of the relationship between P and f

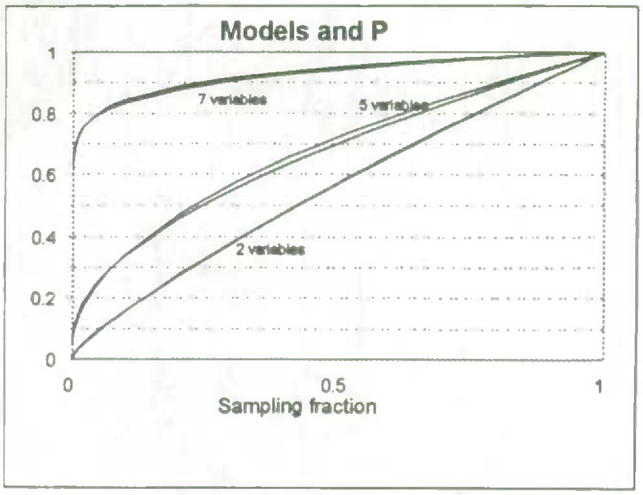
The formulation of this relationship, which very closely corresponds to what is observed, is given by the following expression:

$$P_M = \left| \frac{f + \gamma}{1 + \gamma} \right|^{\alpha},$$

where $0 < \alpha < 1$ and $\gamma > 0$. The parameter α directly influences the observed concavity of the relationship; γ is directly linked to the rate of uniqueness in the population. For the population and the contents given in examples in the preceding section, we obtain values for the parameters α and γ. These values were obtained by using the least squares method.

Content	Alpha	Gamma
2 variables	0.805167	0.002085
3 variables	0.643916	0.002484
4 variables	0.548488	0.001771
5 variables	0.471880	0.003014
7 variables	0.075045	0.001915

The following graphs show that the model cited can be adjusted fairly well to take account of different contents.



We cannot directly use the relationship between P and f to estimate the parameters α and γ, since the conditional probability is not observable. The only quantities of interest that are observable are the components of the sample structure. Let us try to derive the expectation of the number of unique elements in a sample from P and f.

Theorem D: *If the parametric formulation between P and f is correct with parameters α and γ, then the expectation, within the meaning of the model, of the number of unique elements as a proportion of the sample is given by*

$$Q_n = E_m \{ u_1 / n \} = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha},$$

where β is the reciprocal of the multiplication of γ by the size of the population.

Demonstration: By definition, the conditional probability sought is the quotient of the proportions of unique elements in the population and in the sample respectively. Since the formulation between P and f is correct, we have

$$P = \left(\frac{\frac{n}{N} + \gamma}{1 + \gamma} \right)^{\alpha} = \left(\frac{1 + \frac{n}{\gamma N}}{1 + \frac{N}{\gamma N}} \right)^{\alpha}$$

$$= \left(\frac{1 + \beta n}{1 + \beta N} \right)^{\alpha} = \frac{Q_N}{Q_n}.$$

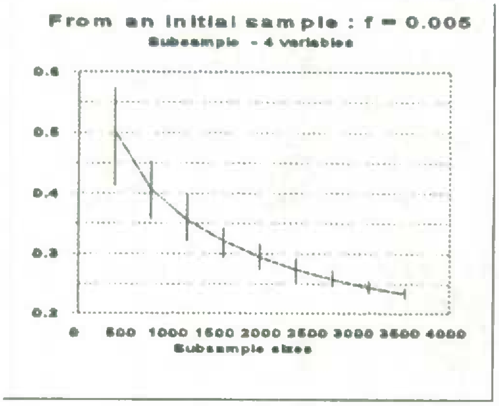
Which gives the following:

$$Q_n = K\left(\frac{1}{1+\beta n}\right)^{\alpha}.$$

Since $Q_1 = 1$, we obtain the desired result.

This theorem tells us why the Poisson-Gamma model has such a great lack of adjustment. That model implies a convex relationship between P and f. But we know by observation that the relationship is concave. Thus, trying to adjust the Poisson-Gamma model to reality can yield something approaching linearity ($\alpha = 0$). This is exactly what the literature tells us.

The question arises as to whether the relationship between Q_n and n corresponds to reality. We have taken as an example the population of size 800,000 with a content of four variables, from which we selected a first sample with a sampling fraction of 0.005. From this sample we selected, completely independently, 900 samples: 100 samples with a sampling fraction of 0.1; 100 with a ratio of 0.2; ... ; 100 samples with a ratio of 0.9. Since the first sample and the others are drawn according to the simple random design, all these samples are drawn according to a simple random design (only the sampling fraction changes). The following graph gives us the empirical curve of the relationship between Q_n and n:



The vertical lines give an indication of the variability of the proportion of unique elements in the samples. The distribution governing the differences is extremely hard to model. On the other hand, the curve of the means of these proportions closely reflects the model.

5.2 Estimation of the parameters α and γ

The model presented in the preceding subsection seems to correspond closely to reality. All that remains is to find a method of estimating the parameters. The model is written as follows:

$$u_1/n = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha} + \epsilon ,$$

where ϵ is governed by the distribution of the differences of the u_1 values. The standard methods of estimating such parameters depend heavily on this distribution. Since we do not know it and we are not in a position to make hypotheses, we will instead use realizations of the means Q_n and assume that these points will be near the curve if the means are based on a number of experiments (e.g., 100 samples). The method is as follows:

I. Select repeated simple random samples from the original sample according to several sampling fractions (e.g. 0.1, 0.2, ... , 0.9). There must be a large number of repetitions for each of these sampling fractions.

II. For each of the sampling fractions, calculate the means of the number of unique elements in the sample (Q_n).

III. Use a numerical method[2] to determine the parameters α and β that correspond the most closely to what is observed.

IV. Determine γ from β.

V. Calculate the conditional probabilities using the model.

With the data from previous examples, if the original sample has a sampling fraction of 0.005, we find the following estimates for the conditional probability:

[2] We used the NEWTON algorithm programmed into the NLIN procedure in the SAS software (version 6.10).

	Alpha		Gamma		Conditional probability	
Content	True value	Estimate	True value	Estimate*	True value	Estimate
2 variables	0.805167	0.538648	0.002085	0.0000222	0.0219	0.0577
3 variables	0.643916	0.422989	0.002484	0.0000981	0.0501	0.1072
4 variables	0.548488	0.374097	0.001771	0.0000961	0.0744	0.1387
5 variables	0.471880	0.274312	0.003014	0.0001914	0.1146	0.2361
7 variables	0.075045	0.068793	0.001915	0.0011052	0.6949	0.7040

* Even if the estimates are close to zero, they are nevertheless significantly different from zero, to judge from the diagnoses produced by the NEWTON algorithm.

Two criticisms of this way of estimating should be noted at the outset. First, there is nothing in the method that can prevent variations in the proportion of unique elements in the sample. Perhaps it would be wise to use a variant of the least squares method by weighting the distance between the model and the points obtained through observation. Second, and even more serious, is the extrapolation required in order to estimate the conditional probability. The problem is that we have to establish the ratio between Q_N and Q_n, and estimating Q_N amounts to extrapolating unduly. On the other hand the method systematically underestimates the parameters and overestimates the conditional probability. Clearly, this method has major flaws from several standpoints, but before casting it aside, we should examine whether the overestimation of probabilities is fortuitous or systematic for real populations. For if it is not fortuitous, then the estimation method is permissible, since it would yield conservative evaluations of disclosure risk. Let us make the following conjecture:

Conjecture E: *The proposed estimation method always yields an overestimate of the conditional probability for real populations.*

The rest of this section assumes that this conjecture is true.

5.3 Rules of confidentiality

A rule of confidentiality is defined as any measure applied to data in order to reduced the risk of disclosure. In the case of discrete microdata, such measures are divided into two groups: suppression, and introduction of noise into the data. A rule is said to be "global" if it applies to all records; otherwise it is said to be "local". Suppression of a piece of data means that it is recoded by use of a new category signifying "suppression for purposes of confidentiality". The elimination of a record or the suppression of one or more pieces of data for this record is known as local suppression. Global suppression is the elimination of a variable from the file. Since disclosure risk and the conditional probability are closely related, the rules of confidentiality must reduce this probability as much as possible.

Looking at the relationship between P and f, we can make the following observations:

I. For a given content, the most secure file, that is, the one that yields the smallest conditional probability, is the one in which the sample size is unity.

II. Even for very small sampling fractions, files that have "explosive" contents may have very high disclosure risks.

An example of an explosive content is one given by seven variables. The last two variables, the sixth and seventh, are respectively the four-digit industry and occupation codes for individuals in the labour force. Of course, it is not possible to have this level of detail in a secure file. On the other hand, it may very well be that for longitudinal files, for example, the risk of disclosure of each cross-sectional part is acceptable but that the risk for the file as a whole is unacceptable. Take as an example the marital status in Canada of an individual. This variable has five values: single, married, separated, divorced and widowed. In and of themselves, these five values may not make this variable very "dangerous," but if the variable is included in a longitudinal survey over ten years, the information provided by this variable for purposes of evaluating risk must also contain all the changes between the values of this variable over time. Thus, instead of having a five-value variable, we must instead consider a variable representing changes from one marital status to another. This new variable may contain several hundred values, with the result that the file content explodes.

Thus small sampling fractions and non-explosive contents guarantee low conditional probabilities. If the model and the conjecture hold up, a low proportion of unique elements in the sample and an estimate of α that is not too low are sufficient to guarantee non-explosive content. The following theorem enables us to determine an upper limit for the proportion of unique elements in the sample.

Theorem F: *According to the model, we have*

$$P = \left(\frac{f}{1 - Q_n^{1/\alpha}\,(1 - f)} \right)^{\alpha},$$

Thus, if the conjecture is true, by specifying a conditional probability P^* not to be exceeded, the following expression

$$\left(\frac{1 - \frac{f}{(P^*)^{1/\hat{\alpha}}}}{1 - f} \right)^{\hat{\alpha}},$$

in which $\hat{\alpha}$ is the estimate of α, gives an upper limit for the expectation of the proportion of unique elements in the sample.

Demonstration. The first identity takes the following form.

$$Q_n = \left(\frac{1+\beta}{1+\beta n} \right)^{\alpha} = \left(\frac{1+\frac{1}{\gamma N}}{1+\frac{n}{\gamma N}} \right)^{\alpha}$$

$$= \left(\frac{\gamma + \frac{1}{N}}{\gamma + f} \right)^{\alpha} \approx \left(\frac{\gamma}{\gamma + f} \right)^{\alpha}.$$

We thus obtain the following expression for γ,

$$\gamma = \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}.$$

If we replace γ by this expression in the definition of the model, we obtain

$$P = \left(\frac{f + \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}}{1 + \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}} \right)^{\alpha} = \left(\frac{f}{1 - Q_n^{1/\alpha}\,(1-f)} \right)^{\alpha}.$$

As to the upper limit, it should first be noted that the function, for x and y between 0 and 1,

$$F(x,y) = \left(\frac{f}{1 - x^{1/y}\,(1 - f)} \right)^{y}$$

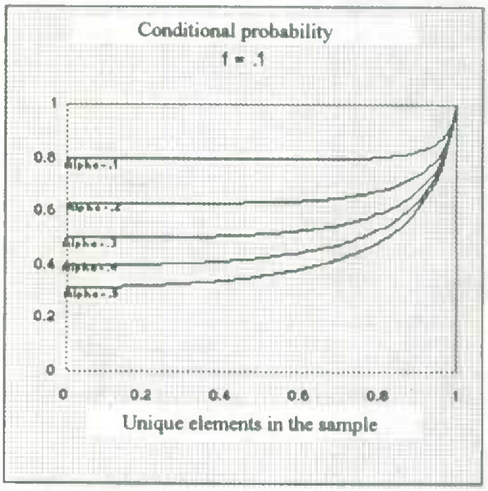
is increasing for x (with y held constant) and decreasing for y (with x held constant). Thus we have

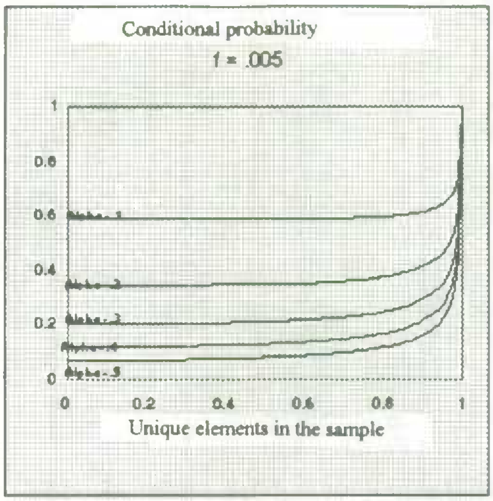
$$P = \left(\frac{f}{1-Q_n^{1/\alpha}\,(1-f)} \right)^{\alpha} \le \left(\frac{f}{1-Q_n^{1/\hat{\alpha}}\,(1-f)} \right)^{\hat{\alpha}}$$

$$\le \left(\frac{f}{1 - \left(\frac{1 - \frac{f}{(P^*)^{1/\hat{\alpha}}}}{1 - f} \right)(1 - f)} \right)^{\hat{\alpha}} = P^*.$$

The first inequality comes from the assumed underestimate of α, and the second from the application of the limit for Q_n.

The following graphs give the relationship between P and Q_n for different contents and sampling fractions.





These graphs clearly show us that we can tolerate a great number of unique elements in the sample for small sampling fractions. But what should be done if the proportion of unique elements in the sample is too high? The following section will seek to answer this question.

6. PROCESSING THE DATA

The preceding section offered ways to assess disclosure risk. Here again, sufficient conditions for having a low risk are:

I. A small sampling fraction;
II. A fairly low proportion of unique elements in the sample;
III. An estimate of α that is not too small.

If any one of these conditions is not met, it is necessary either to sample again in order to reduce the sampling fraction or change the content. Any change in content will be referred to as processing the data. This may be either global processing or local processing. Global processing is processing done to all the records, such as in combining values of one of the matching variables. Local processing, on the other hand, is done to only some of the records. There are several methods of global or local processing. They all have their strong points and weak points. The objective of this section is not to go over the list of methods and describe the performance of each. Rather, we would like to answer the following question: when local processing is decided upon, which are the records that should be processed in order to optimize the operation?

In theory, the purpose of the processing is to reduce the number of unique elements in the population that are contained in the sample. Hence if we want to optimize, we must find a means of identifying these records and then process them in such a way that they are no longer unique in the population. As regards the processing to be carried out, we can assume that the necessary changes to certain values of the matching variables are minimal. We will therefore assume that once a record is processed, it will be considered secure. What remains, then, is the matter of choosing the records to be processed. Since the unique elements in the population are necessarily unique in the sample, we should first concentrate solely on the unique elements in the sample. But this is not enough. We must be able to screen the unique elements in the population from those that are unique only in the sample. It is here that the concept of the multiplicity of a record comes into play.

How can the unique elements in the population be screened from the others? It is necessary to be able to evaluate the "degree of uniqueness" of the records -- to be able to say that one record is more unique than another. How is this to be done? We shall formulate a postulate and see where it leads us.

Postulate G: *Most of the unique elements in the population are also unique in the population for a limited subset of matching variables.*

This postulate states that the attribute of uniqueness in the population mainly depends on a very rare combination of values of a small number of matching variables. This being said, if we try to identify unique elements in the population with, say, only three matching variables, perhaps some elements will already be classified as unique. But by looking for unique elements for all combinations of three variables from among all matching variables and by adding together, for each element, the number of times that it is unique, we arrive at a quantitative notion of uniqueness. The number of times that an element is unique in a three-dimensional table is called the multiplicity of this element. We can say that the higher the multiplicity of an element, the greater the risk of its being identified. What happens when we have only a sample? We found that if we calculate multiplicity using only the sample, it defines a partition of the sample, the different parts of which have very different proportions of elements unique in the population. We simulated a small example in order to show the usefulness of multiplicity.

We took a simple random sample with a sampling fraction of 0.009 from a population of 781,825 elements. This yields a sample size of 7,037. The file contains five matching variables. The number of unique elements in the population is 35,718 (4.5%). The number of unique elements in the sample is 2,301

(32.7%). The number of dangerous elements (those which are unique in the population and included in the sample) is 321 (4.5%). The conditional probability is 14%. If we choose randomly from among the unique elements in the sample, only 14% of these records (on average) are dangerous. Thus considerable processing is carried out on records that do not require it. If we calculate the multiplicity of the records, we obtain the following table:

Results of simulation

Multiplicity	# elements	# uniques	%
10	18	15	83.3
9	41	23	56.1
8	64	33	51.6
7	45	26	57.8
6	191	61	31.9
5	220	77	35.0
4	140	33	23.5
3	388	32	8.2
2	294	17	5.8
1	472	3	0.6
0	5,164	1	0.0
Total	7,037	321	4.5

We can easily see that the partition generated by multiplicity greatly helps us in choosing the records to be processed. For example, if we decide to process all records having a multiplicity greater than three, we eliminate 83.4% (268 elements) of the dangerous records by processing only 10.3% of the records, which is more effective than going about it randomly. We tried this technique with files containing ten or fifteen matching variables, and while the screen was not as effective as the one shown above, the results are nevertheless surprising. Research is now under way to determine the minimum multiplicity at which processing would be necessary. This multiplicity, called the "singularity threshold," would indicate, if the processing percentage is too high, that more comprehensive measures need to be considered.

7. CONCLUSION

In this article we have described the state of research at Statistics Canada on evaluating disclosure risk in microdata files containing discrete variables. In order for the model described in this article to stand out from the others, it must be possible to justify it theoretically. The research begun by Skinner on the Poisson-Lognormal model is encouraging. If these models correspond to reality, they should converge at some point. Also, the multiplicity of records is a concept which in our view is essential for the efficient processing of the microdata file.

REFERENCES

Greenberg, B. V., and Zayatz L. (1992). Strategies for Measuring Risk in Public Use Microdata Files, Statistica Neerlandica.

Bethleem, J. G., Keller, W. J., and Pannekoek, J., (1990). Disclosure Control of Microdata, *JASA*, 85, 38-45.

Skinner, C. J., and Holmes, D. J. (1992). Modelling Population Uniqueness, Paper presented at International Seminar on Statistical Confidentiality, Dublin.

Goodman, L. A.(1949). On the Estimation of the Number of Classes in a Population, *AMS*, 20, 572-579.

Mercier, A. (1984). Quelques identités de l'analyse combinatoire, *Discrete Mathematics*, North Holland, 49, 139- 149.

SESSION 6

Making Data Accessible to the General Public

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

MAKING INFORMATION AVAILABLE FOR MARKET RESEARCH AND SPATIAL ANALYSIS

C. Sewards and L. Li[1]

ABSTRACT

The prototype version of the Market Research Handbook provides for user-friendly access to a broad range of integrated Statistics Canada data contained in the Market Research Handbook, facilitating computer-based analysis of products and markets. This prototype is an integrated system which utilizes the functionality of three types of software: hypertext, spreadsheet, and mapping, and is fully compatible with popular word processing and spreadsheet softwares. This product takes the form of an electronic book with a searchable table of contents linked to tables within tables in the spreadsheet and maps in the geographic viewing engine. This suite of software enables users to perform statistical analysis as well as displaying data in graphic or map format. This integrated method of delivery serves as a potential low cost dissemination vehicle for a wide variety of information.

KEY WORDS: Market research; Geographic information systems; Spatial Analysis; Data Integration.

1. INTRODUCTION

The Market Research Handbook (MRH) is one of the "flagship" publications of Statistics Canada. It contains data on a wide variety of subjects, from consumer purchasing patterns to international trade, the socio-economic character of different Canadian cities, etc., at highly aggregate levels. As such, it provides users with insights on broad market conditions in the Canadian marketplace. As well, because it includes data from almost every subject matter area within Statistics Canada, it is a good tool for leading analysts in their prospecting for data to support their market research needs.

Recently, a pilot study was completed to examine the conversion of the MRH from a print product to an electronic product. The results demonstrated the potential of an electronic delivery vehicle, which takes advantage of the strengths of hypertext software, a spreadsheet and a Geographic Information System (GIS), to provide users with a powerful and flexible product meeting many of their business requirements. At the same time, this enabling technology raises the issue of data harmonization to the fore.

This paper presents the results of the pilot study. The discussion begins with a look at user needs which were identified through client consultation and focus groups. Technical issues and data integration issues which were encountered in the development of the electronic prototype are then examined. Details of the solution, using a combination of hypertext, spreadsheet and Geographic Information System (GIS) softwares to deliver a series of integrated data sets from many subject matter areas within Statistics Canada are then detailed. The paper concludes with some thoughts on the potential of this approach for the future.

2. USER NEEDS

The MRH serves an extremely diverse client base. After review of the client feedback and examination of the desired focus of the product, the core applications for the product, for the purposes of this study, were considered to be:

- examine market conditions;
- assess market potential;
- prospect for customers;
- analyze broad market trends.

Focus groups and client feedback indicated that the

[1] Crystal Sewards and Larry Li, Geography Division, Statistics Canada, Ottawa, Ontario K1A 0T6.

MRH was a useful product, however, being over six hundred pages in its hardcopy format, it was straight-jacketed by the inflexibility of the paper medium. Clients commented on the difficulties in finding the data they needed, the lack of detail in the data, and wished for an electronic version of the product with some data manipulation and analytical functionality. They wanted: more data (more longitudinal data, more detailed geographic breakdowns, and more detailed commodity categorizations), more flexible data, more detailed data, more source notes, better references to complementary data sources, the ability to be able to generate cross-tabulations, analytical measures, charts, tables and maps, and to export data to spreadsheets or other software for inclusion in written reports. They wanted a premium product that met their needs, and were willing to pay more for it.

The above client "wants" set the basic specifications for the prototype product. As well, the MRH management team wanted to maintain strong ties between the electronic product and the current publication. They felt that this would enable them to take an evolutionary approach to the growth of the electronic product, as it was unlikely that the first electronic version could include all the data which the users desired. Further, to guide product development, a target price of $500 was set for the proposed product. This would put the product squarely in the range of similar products currently in the marketplace. Of this amount, a limit of $100 was set for software.

3. THE SEARCH FOR A SOLUTION

The task began with a review of potential software which would be able to deliver the required functionality. Four types of packages were considered: hypertext packages, spreadsheets, database packages and low-end GISs.

Hypertext software, such as Folio Views, has excellent facilities to model information in the form of an electronic book. It is adept at accepting inputs from common word processing packages to create a text database on which keyword searches can be executed. Many have the ability to imbed "jump links" to allow a reader to quickly skip from one section to another. These links can also be used to associate footnotes and other explanatory comments to selected words or phrases. However, hypertext packages are less adept at handling data tables. They generally lack facilities for mapping and graphing data or for statistical data manipulation.

Spreadsheets are extremely well suited to handling numeric tables. Many include powerful mathematical and statistical functions. Graph and charts can be generated at the touch of a button. Rudimentary mapping capabilities are also emerging in some popular packages. Some specialized spreadsheet-type software, such as Ivision are able to handle very large tables. They can reorganize the position of variables within a table; rearrange the X and Y axis of a table and change the "dimension" being displayed very efficiently. On the down side, they are less adept (than other hypertext packages) in the handling of formatted text data; performing text oriented queries, and associating explanations with keywords or phrases.

Database packages, such as dBASE or FoxPro, handle numeric data and relatively short text string very well. Powerful operators are available for indexing and searching of information, especially within given fields. They are able to execute operations to combine data from different fields and create joins between tables. However, in general, these packages are not primarily aimed at text handling or mapping.

Geographic Information Systems (GISs) commonly incorporate a database engine tied to a mapping engine. The former providing the facilitates for handling attribute data as previously discussed in the section on database packages, while the latter provides the functions for mapping, spatial queries and spatial data integration by overlaying different maps or layers of data. As in database packages, GISs do not handle text descriptions well. Statistical operators are also often missing.

At the conclusion of the software review, it was apparent that no single software available at the time could fulfill all the user requirements without significant custom programming. Fortunately, with the, then, recent release of ARCView-1 into the "freeware" (but still licensed) domain, and favourable licensing terms for both Folio Views and Ivision for mass distribution, it appeared that integration of the three packages into a single delivery vehicle for the electronic MRH could be considered within the target cost envelop. Thus, the decision was made to develop the prototype using all three packages in an integrated manner. In this way, the strengths of each package could be used to meet different requirements on the specification list.

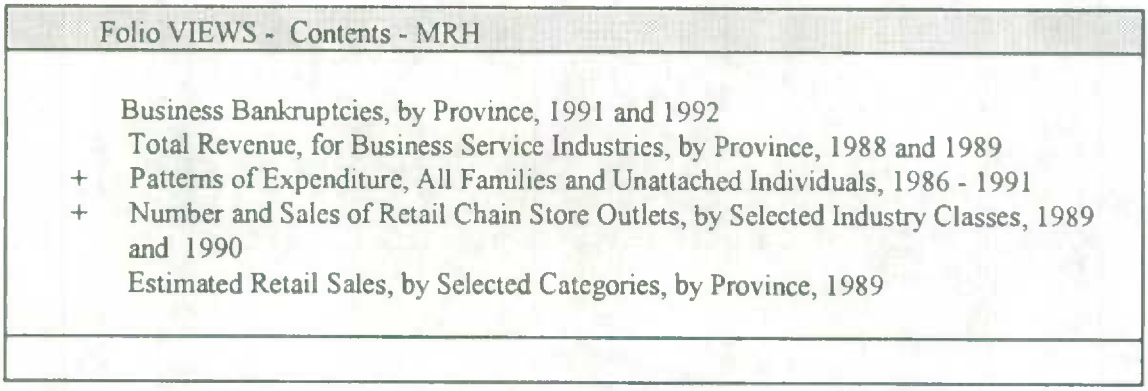
4. THE INTEGRATED PUBLISHING VEHICLE

The integrated publishing vehicle for the prototype product is constructed very much like an electronic book.

It uses the strength of hypertext to provide the intuitive structure of a book, but takes advantage of the advanced navigation tools of the software to make finding information easier, as well as providing context sensitive notes. To deliver the table handling and data manipulation capabilities which were desired, all the tables were carried inside the spreadsheet engine, in this case Ivision. Mapping and spatial analysis of marketing information was desired by users. The spatial view of information was considered to be most valuable in dealing with information at a subprovincial level where many cells of data would be involved, and where the location of customers or competitors were of interest. These capabilities were delivered by incorporating a Geographic Information System (GIS), ARCView-1, into the publishing vehicle.

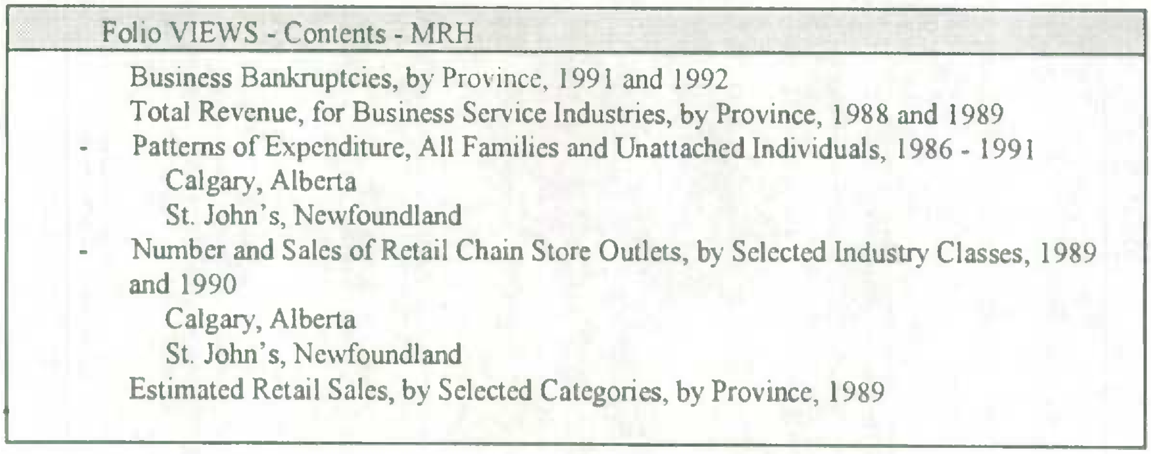
Although the structure of the integrated publishing vehicle is of interest from a cost and technical standpoint, users perception of its user friendliness and functionality is perhaps even more important. In the prototype, the user enters the electronic MRH via Folio Views. The first screen the user sees is the Contents Window which contains a listing or table of contents for all tables contained in the Folio database/Infobase.

This table of contents is expandable and collapsable, allowing the user to see as much or as little information as needed. A plus sign (+) next to an entry indicates that the entry can be expanded to show subordinate levels. A minus sign (-) indicates that the entry cannot be expanded further, but may be collapsed. Access to any table in the list is gained simply by double clicking on it; each entry in the table of contents is linked automatically to the appropriate section in the document. Beneath each table title are two icons: one Ivision icon and one ArcView-1 icon. Double clicking on either executes the given software and opens the data table, in the case of Ivision, or the map view, in the case of ArcView-1. Closing either Ivision or Arcview-1 after viewing a data table or a map immediately sends the user back to Folio Views at the original table title. Sources and footnotes for each table can also be accessed by double clicking on the pop-up notes beneath each table title in Folio Views. Once a user has moved from a Folio table title to the data table within Ivision he or she has the ability to sort the data for a given variable in ascending or descending order, change the position of any dimension, nest dimensions, or even calculate percentage change between years of data with just a few mouse clicks. This data can then be quickly graphed to highlight changes and facilitate comparative analysis. Then by clicking on the ArcView-1 icon the user can map the data to bring out spatial patterns.



Folio VIEWS - Contents - MRH

Business Bankruptcies, by Province, 1991 and 1992
Total Revenue, for Business Service Industries, by Province, 1988 and 1989
+ Patterns of Expenditure, All Families and Unattached Individuals, 1986 - 1991
+ Number and Sales of Retail Chain Store Outlets, by Selected Industry Classes, 1989 and 1990
Estimated Retail Sales, by Selected Categories, by Province, 1989

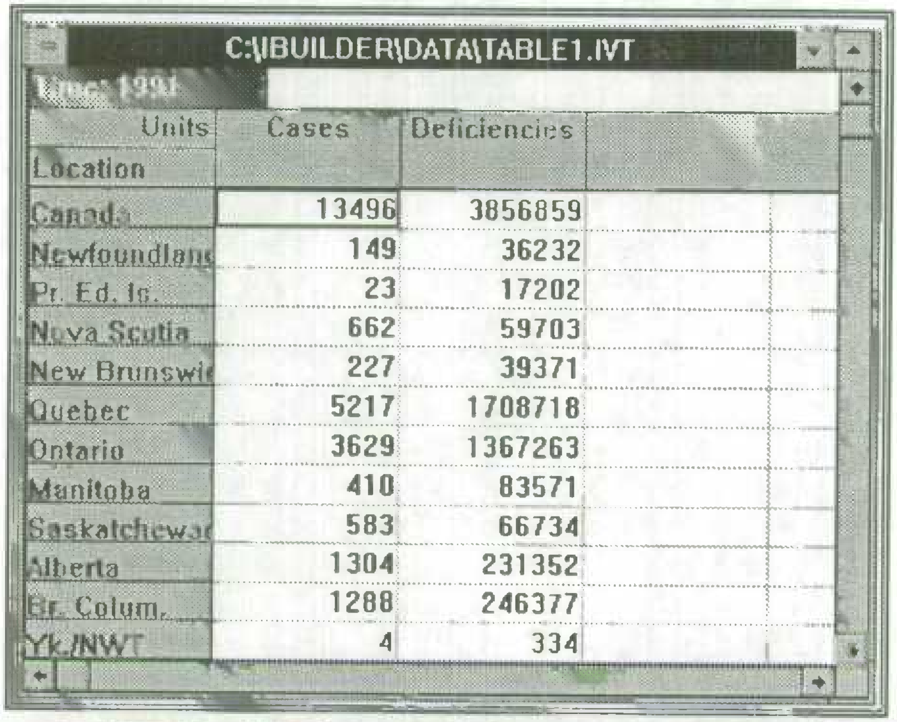


Folio VIEWS - Contents - MRH

Business Bankruptcies, by Province, 1991 and 1992
Total Revenue, for Business Service Industries, by Province, 1988 and 1989
- Patterns of Expenditure, All Families and Unattached Individuals, 1986 - 1991
Calgary, Alberta
St. John's, Newfoundland
- Number and Sales of Retail Chain Store Outlets, by Selected Industry Classes, 1989 and 1990
Calgary, Alberta
St. John's, Newfoundland
Estimated Retail Sales, by Selected Categories, by Province, 1989

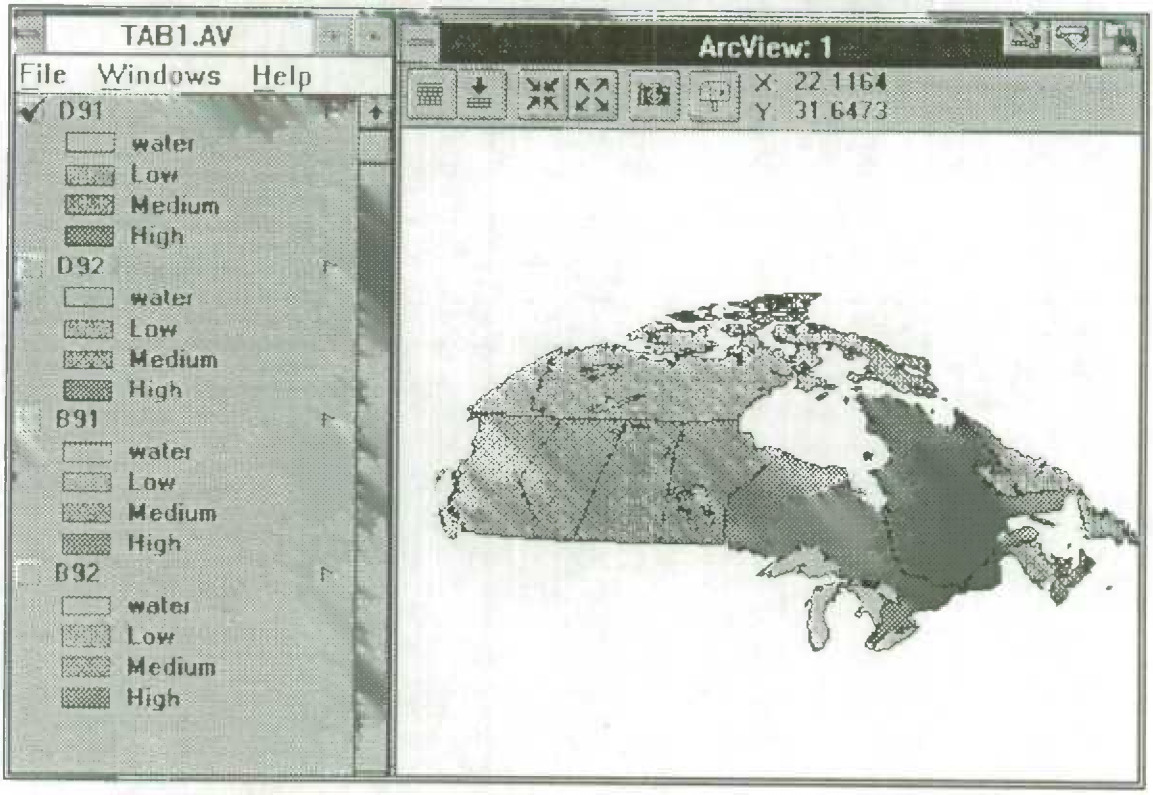
The Query option available in Folio Views allows the user to search the Folio Views database for a given key word such as bankruptcies or Calgary. A list of the number of occurrences of each key word appears in the query window and by clicking 'apply to all' the user is immediately taken to the first occurrence of the word in the infobase. Clicking the next button takes the user to subsequent occurrences of the key word.

ARCView-1 provides the mapping and spatial analysis functions. It allows the user to display the data within a given table in the form of a map. The user can pan, zoom in or out, and view the data table for each map. For the advanced users, the choice to drop into ARCView-1 directly to access much fuller functionality is also available, since the full software is included.



Time: 1991

Units Location	Cases	Deficiencies
Canada	13496	3856859
Newfoundlan	149	36232
Pr. Ed. Is.	23	17202
Nova Scotia	662	59703
New Brunswi	227	39371
Quebec	5217	1708718
Ontario	3629	1367263
Manitoba	410	83571
Saskatchewa	583	66734
Alberta	1304	231352
Br. Colum.	1288	246377
Yk./NWT	4	334



X: 22.1164
Y: 31.6473

5. DATA ORGANIZATION

Driven by the desire to maintain a strong link between the electronic and print product, the prototype began by retaining the structure and basic table organization of the current book. A set of seven tables were selected from the current book and implemented in the prototype in order to demonstrate the functionality of the electronic product. The prototype gave users the ability to examine the functionality of the concept using realistic queries. These queries included some basic market research questions: Who are my customers?; Where are my customers?; and How do I find more like them?

In addition, the prototype demonstrated the technical feasibility of integrating more data and more detailed data into the product. This was demonstrated by the inclusion of more years of data for time series inside the Ivision spreadsheet. For increased geographic detail, maps were attached to provincial data variables, providing a subprovincial view of the topic. With the subprovincial maps, users can access the associated data in the database.

As prototype development progressed, it became apparent that the current structure limited the ability of the user to fully take advantage of some of the opportunities offered by the softwares to search, integrate and sort. One of the main barriers is the storage of data in many, small, separate tables in the current product. The highly partitioned structure limits the ability of search facilities as they operate only within a given database. Data manipulation across tables is also cumbersome, requiring users to open-up and clip pieces from different tables to do analysis. Thus, it may be desirable to combine tables together into a larger database.

Although the scope of this study precluded in-depth examination of these issues, preliminary results indicated that integration of much more data into a single database may enable users to quickly resort data and display all the cells of data for the given place or time frame in one table, thus facilitating easier analysis of trends and market conditions. This holds promise for better meeting a number of user requirements; however, they are not without challenges. Some of the challenges and issues which came to the forefront when data reorganization was attempted included data availability and data compatibility. For example could a table for population by province for 1986 and 1991 be integrated with a table on household expenditures by province for 1990? Or could a table for residential construction starts by province for 1992 be integrated with a table on household income by census metropolitan area for 1992? Should the data be grouped by geography, subject matter, year, or some other variable? Some data collected for a given geography was collected in different years; how could these tables be integrated or cross tabulated? Users also suggested providing some data at more detailed geographic levels; certain data is available at higher geographic levels such as provinces or census metropolitan areas (CMAs) but is not available at the census subdivision (CSD) level. Would it be acceptable to provide different or less comprehensive data at these lower geographic levels? What other layers of data are available for the same geographies already presented in the MRH? Presentation of the data also became an issue; with an electronic version perhaps data could be disseminated for only one province or one CMA. These and other questions will require more careful examination in the future.

6. CONCLUSIONS

The results demonstrate the potential of an electronic delivery vehicle, which takes advantage of the strengths of hypertext software, a spreadsheet and a Geographic Information System (GIS), to provide users with a powerful and flexible product meeting many of their business requirements. Such a vehicle can deliver ease of use; carry much more data; and enable users to manipulate the data to suit their analytical needs; all within an attractive price.

At the same time, this enabling technology points out some of the conceptual differences which exist within variable definitions for different Statistics Canada data sets. For example, the definition of 'family' varies between tables. The Survey of Family Expenditures defines a 'family' as a group of people living in the same dwelling who depend on a common or pooled income for major expenses. For tables which relate to Census data however, a family is comprised of several people living in the same dwelling who are husband and wife or common-law partners, with or without never-married sons or daughters at home, or a lone parent and at least one son or daughter who has never been married.

The attempt to create an integrated electronic product also brought to the forefront the issues of data harmonization and price and product line rationalization. Efforts are already underway to examine aspects of these challenges, which in the end will enable Statistics Canada to better meet users' needs.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

SEDUCING THE GATEKEEPERS: STATISTICS CANADA'S *DAILY* AND THE NEWS MEDIA

W.R. Smith[1]

ABSTRACT

Statistics Canada aimed a two-year initiative at improving communication with the general public through the news media. Its focus on exploiting the media as the port of dissemination required changing the corporate culture and garnering support at the highest levels. Through formal and informal programs, and sometimes through trial and error, fundamental changes have been made in the way the Agency's data releases are written, specifically to include trends and analysis in understandable language and with clear and illustrative tables and charts. These processes have included the deliberation of a senior editorial board, published guidelines, consultation services and media-oriented writing courses. Although success is difficult to measure, to date the initiative has produced increased and higher quality media coverage, as well as added benefits such as more positive and enthusiastic attitudes in dealing with the media.

KEY WORDS: Media; Consulting; Analysis; Editorial Board; Communication; Guidelines.

1. INTRODUCTION

1.1 Serving the Public

For the past two years, Statistics Canada has directed much of its energy into improving communication with the general public through the news media. The initiative reflects an underlying conviction that the general public is an important target audience for a national statistical agency.

Efficient communication with the general public, however, requires the intervention of the news media. Journalists are the gatekeepers between the statistical agency and the public. In a democratic society, journalists are fiercely independent and can be neither bought, commanded nor co-opted. They must be won over—seduced—into acting as conduit for the statistical agency's information.

Statistics Canada's first line of communication with the media is its official release vehicle, *The Daily*. Agency policy requires that the availability of data be officially announced through *The Daily* before being disseminated further. With publication every business day, *The Daily* was the ideal vehicle to improve communication with the broad public through the media.

This paper describes the environment and culture in which the initiative was developed. It describes the objectives and the processes used to achieve them. And finally, it briefly discusses the results to date.

1.2 Statistical Agencies and the General Public

One conviction is central to Statistics Canada's initiative to communicate more effectively through the news media: The general public is an important audience for a statistical agency. In a liberal–democratic society the public, as citizens and economic players, require information on the population, society, economy and culture of the nation. This information will guide them in doing their jobs, raising their families, making purchases, forming opinions of their governments, in voting, and in making myriad daily economic choices. A well-informed population improves both the political and economic efficiency of a nation.

At the same time, the consequences of casting an individual vote, or making a personal economic decision scarcely warrant an extensive search for information. Few individuals will visit their statistical agency to compile a national economic overview before voting. And if they did, agencies could scarcely cope with the demand. Similarly, it is beyond the means of a statistical agency to communicate directly with all citizens every day.

The news media, therefore, afford statistical

[1] Wayne R. Smith, Communications Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

agencies a unique opportunity to fulfil a critical part of their mandate. Citizens glean information from newspaper articles and headlines, and many listen to one or more television or radio broadcasts every day, further shaping their views. The extent to which statistical agencies can gain access to the news media and communicate effectively through them has an enormous impact on how well they can inform the general population.

In addition to the direct benefit of a more informed citizenry, Statistics Canada anticipated a number of secondary benefits. Through a more focused public communications program, the population would become more aware of the statistical agency and the importance and relevance of its programs. Respondents would see the benefits of co-operation with Agency surveys. And the Agency's mainstream clients would become more aware of the information available. In short, Statistics Canada anticipated greater public support, improved respondent relations and marketing benefits.

1.3 The Gatekeepers

In Canada, as in most liberal–democratic countries, the news media fiercely defend their editorial autonomy. Media coverage is obtained by being newsworthy—or perhaps newsworthiest—in relation to competing stories of the day. Journalists provide coverage in their own words, not those of the source, and coverage reflects the ability of the journalist to understand the information. They are very much gatekeepers of access and meaning between statistical agencies and the general public.

Newsworthiness is a relative term. Editorial capacity—the amount of information that can be tucked in between the advertisements—is limited and, in recent years, shrinking. Yet competition for media coverage is growing as more interest groups and organizations vie for attention to their issues and actions through the media. While journalists, in Canada at least, recognize the national statistical agency as a major news source, how much coverage they provide depends largely on how well the news value of a release is conveyed, the number and nature of "competing" news stories, and editorial capacity. In short, if you want coverage, you must show clearly why anyone should care about the information. Coverage is scalable too, with the more newsworthy stories earning more space.

Journalists have learned that *The Daily* provides a comprehensive, one-stop overview of new information available from Statistics Canada. The news media closely monitor *The Daily*. Competition is fierce among the major newswires when *The Daily* is hand-delivered to the Press Gallery in downtown Ottawa at the 8:30 a.m. official release time.

Obtaining coverage is only the first obstacle to effective communication through the news media. The news media, if a mirror, are an imperfect one. Most journalists work to tight deadlines, producing three to five news stories a day. Even if they have the capacity to analyse raw information independently, they lack the time. Other than feature writers, journalists will add little analytical value to the information provided. But increasingly, journalists lack the in-depth training to fully analyse raw information. Reduced editorial budgets have made "beat" journalists, who can build up specialized knowledge, a declining phenomenon. Today's news coverage reflects journalists' understanding of information provided and if it is unclear, the reflection may well be wrong and everyone loses.

So obtaining positive, accurate and informative media coverage in most liberal–democratic societies is an exercise in seduction. We must show *why* our information is important enough to warrant attention. If we want our information to be presented as anything more than factual trivia, we must entice journalists into covering our news by providing analysis in context and showing trends that bring out its true signficance. And we must communicate clearly in simple language if we want our information echoed accurately to the general population.

These activities serve our own requirements and those of the general public, not the needs of the news media. In this symbiotic relationship, the media benefit certainly, but statistical agencies and the public benefit even more.

1.4 Objectives and the Internal Environment

Senior management at Statistics Canada had long felt the Agency could, or should, achieve more in passing information through the news media. Both the quantity and quality of media coverage on the Agency's information releases could be improved. Improving the content of *The Daily*, specifically the major release section, became the focus of the Agency's initiative.

Central to this objective was making fundamental changes in the way information releases were written for publication in *The Daily*. The news media were explicitly defined as the target audience for *The Daily*, while other users of *The Daily*, including mainstream clients, became secondary.

Statistics Canada has, over time, evolved an extremely orderly mechanism for communicating information to the news media. Under Agency policy, all new releases must be made through Statistics Canada's release vehicle, *The Daily*. The most significant releases are presented with substantial write-ups and are termed "major releases"; less significant releases are covered more summarily.

Almost 30 divisions are involved in writing releases for *The Daily*, along with dozens of individual analysts and managers. The text is developed in the various subject-matter divisions with central communications staff providing light editing, formatting, standardization and quality control. Implementing this new vision would require changing the behaviour of all participants. Knowing what needed to be done was not, in itself, sufficient to get it done.

1.5 A New Model

The Agency set out to improve its releases by expanding analysis to show the importance, relevance and context of new information releases using clearer and less technical language, charts and tables. The relationship of data in one release to that of another would be drawn more clearly. If journalists could see the newsworthiness of new releases, and understand them, they would echo them accurately to the general population.

Communications staff consulted with journalists, emphasizing the major newswires and newspaper chains, to determine how releases could better serve them. Journalists assured the Agency they considered Statistics Canada a major, national news source. They stressed the need to identify more clearly the most important story in a release and to support it with enough analysis and context to allow them to build their own stories. They called for brevity, clear language and simplifed texts and charts. Using this feedback, communications staff constructed and distributed a draft guide to writing effective releases.

Authors in subject-matter divisions were asked to identify an explicit story line and illustrate it through analytical and contextual commentary. They were asked to adopt a more journalistic style and structure, and use plainer language and simpler graphics and tables. This was a revolutionary change from the prevailing "elevator analysis"—what's up, what's down—style of release, and of bulleted, unconnected highlights presented in templates used from release to release with standard tables and graphs.

Analysts were frequently enthusiastic about the new direction, although some subject-matter analysts and many middle managers were reluctant to move to the new model, with middle managers and managers being most resistant to change. The conservatism had its roots in fear and resentment of the news media (not entirely unfounded), concern over the reactions of policy departments to more extensive analysis, and an awareness of data users' needs over those of the media. Closer to home, divisions were concerned about the resource and time requirements for the new model, and uneasy over the extent of editing by communications staff.

Historic release practices rooted in quite directive policy statements supported the conservative view. The traditional response to deep-seated concern about precisely where the line between legitimate analytical commentary and inappropriate policy commentary lies was to err substantially on the side of caution.

It became clear early in the process that simply announcing the new approach was insufficient incentive to change. A multi-part plan that addressed issues of corporate culture, incentives and techniques was developed to achieve the necessary culture shift and implement the new release strategy.

2. GETTING THE MESSAGE OUT

2.1 Changes at Several Levels

Achieving this kind of radical change required action on a number of levels. An initial attempt by communications managers to promote the change on its merits alone, by focusing on the analysts who write releases, failed to deal with resistance from the managers overseeing the process. A more ambitious program was introduced.

The first challenge was to alter corporate thinking about releases and garner management support and collaboration. For two years, every opportunity was taken to present the new model and explain the thinking behind it. To launch the process, the plan was discussed at the corporate policy committee comprising the Chief Statistician and Assistant Chief Statisticians. Assistant Chief Statisticians carried the message back to their managers. The Chief Statistician reinforced the message in a memo and in his annual "state of the union" presentation to managers, and in his annual interview in the employee newsletter, *SCAN*. Training courses on analysis incorporated the message into the course presentations; in fact, Statistics Canada's principal course on data analysis and presentation includes some two days of presentations on the media and release writing. Leading journalists were invited to speak on reporters' working conditions and requirements. The message was pervasive and signalled clearly that change in *Daily* releases was a priority supported by senior management.

2.2 The Senior Editorial Board

Many priorities compete for managers' attention in a large organization. A Senior Editorial Board for *The Daily*, chaired by the Chief Statistician with Assistant Chief Statisticians and senior analysts as members, was established to place this priority at the top of managers' agendas and hold it there. The Board met weekly to review *Daily* releases and to distill guidelines on writing effective releases from them that would guide managers and analysts. All major releases were assigned to a Board member for critical review. Managers and analysts from the subject-matter division responsible for the release were invited to participate in the Board's discussion of the review. If the Board felt the release had been sufficiently improved after several reviews, it was exempted from further scrutiny. To be "graduated" from the review process was a sought-after status. If the release still fell short of the mark, reviews continued. It was 12 months before all major releases were exempted.

Creating the Senior Editorial Board proved the key to the success. The Board's widespread influence and involvement at the corporate level legitimized the process needed for a change of culture. The guidelines on effective release writing—a by-product of the Board's process—were also legitimized. It provided and maintained a high level of energy and focus until results had been achieved. Only by complying with the review process could the managers end the review cycle—a strong incentive to take action and achieve results. Communications managers at lower levels in the organization could not achieve the legitimacy or the focus necessary to bring about radical change.

2.3 Direct Assistance

While the information campaign and Senior Editorial Board created the climate, legitimacy and incentives to improve *The Daily*'s releases, subject-matter divisions required more direct assistance in making the transition.

To a significant extent, the Board itself provided this assistance through the detailed guidance and explicit suggestions for revamping releases that characterized its reviews and discussions. And, by distilling and publishing general guidelines on how to write effective releases, the Board helped divisions make the transition. Still, additional, practical assistance outside the relatively charged atmosphere of the Board meetings was required. This practical assistance was provided in three ways.

2.3.1 Working Groups

In the early stages of the Board's work, relatively formal working groups comprising senior analysts, communications staff and staff from the responsible subject-matter divisions were established for several releases. These mini task forces reported their conclusions along with revised models of the releases to the Senior Editorial Board. This technique helped the Board quickly identify and begin applying broad principles of what makes a release effective.

2.3.2 Consulting Service

The second technique for assisting divisions was to create a consulting service within the Communications Division, linked to the *Daily* editorial staff and available to any subject-matter division. Communications Division brought on staff a journalist who had recently headed the Ottawa bureau for a major Canadian daily newspaper and teamed him with a junior analyst. The team approach balanced the journalist's strong media orientation with the analyst's awareness of analytical issues, data limitations and the boundaries of legitimate comment by a statistical agency. Both participated in all Board meetings and developed the guidelines based on the Board's discussions. Many divisions took advantage of the service, which is now a permanent fixture of the communications program. Some managers identified this service as the most useful help available in redeveloping their releases.

2.3.3 Formal Training Courses

The final form of assistance for subject-matter divisions was a formal training course called "Writing Effective Releases for *The Daily*." The course, developed around the Board's guidelines, included information about the initiative's importance, sensitivity to the working conditions and constraints of the news media, and specific exercises on writing in a journalistic style. The course is offered on a division-by-division basis and is directed at all analysts who write for *The Daily*. Managers responsible for approving releases for publication are asked to attend with their staff since a frequent issue raised by analysts is that managers are reluctant to approve releases written in the new style. Senior Editorial Board members are asked to present the opening remarks and the instructors are, wherever possible, the communications staff and analysts subject-matter staff will in fact be working with during the release process. In addition to teaching new techniques, the course offers a way to establish a network of contacts and a climate of confidence among participants and instructors. Once all interested divisions have participated, it will be offered as part of the standard course offerings.

3. CURRENT STATUS

3.1 Suspension of the Board

In September 1995, with most releases exempted from the Board's critical eye, the Senior Editorial Board suspended its reviews. Although the Board recognized that there had been some backsliding, it concluded that continued review of releases was unlikely to produce further, significant improvements in the short-term. However, the choice of the term "suspended" was deliberate, and a return to the review process is not precluded. The Chief Statistician continues to provide specific feedback on releases to subject-matter divisions when warranted.

The release consulting service in Communications Division, together with the training course still cycling through the subject-matter divisions, continues to drive the process of revamping *The Daily*.

4. RESULTS

4.1 Objectives Largely Achieved

In suspending its review process, the Senior Editorial Board clearly communicated that its objectives had been largely achieved. A comparison of *Daily* releases from the period before the initiative with those of today illustrates the dramatic change.

Feedback from the news media is extremely positive. Interestingly, journalists—whether they are feature writers specializing in economics or social affairs, "beat" reporters, or general news reporters—all say that the strong story lines, analytical background and clearer presentation are helpful in their work. Many general readers of *The Daily* have also commented that they find the revamped releases more accessible, more interesting and more illuminating. This is particularly encouraging as electronic dissemination of *The Daily* allows us to reach a much larger audience. Nonetheless, Statistics Canada has concluded that *The Daily* cannot effectively serve two masters. Concerns by subject-matter divisions that the new style of release writing would not satisfy their mainstream clients have been partially confirmed. While the media permit us to reach millions of Canadians, the direct readership of *The Daily* is limited to a few hundred. The news media will remain *The Daily*'s focus. Subject-matter divisions have been asked to explore new mechanisms, such as day-of-release fax services, to meet other clients' needs.

4.2 Measuring Success

The effort to improve the effectiveness of *The Daily* has been an unmitigated success. Quantitative evidence of expanded media coverage, such as increased lineage or greater numbers of media reports on Agency releases, is elusive. Too many factors determine coverage to permit an aggregate analysis. For example, a poorly written release of two years ago announcing a startling development would inevitably garner more media coverage than a well-written release today announcing no significant developments. A poorly written release may generate more coverage than a well-executed release, but the coverage will be damaging. Coverage of identical releases at two different points in time will be affected by competing, or complementary news events. And, of course, coverage for identical releases would be affected by the media relations effort by communications staff. Aggregate, quantitative analysis of the initiative's impact on media coverage is only possible by controlling these extraneous variables.

4.3 Case Studies

Case studies of individual releases suffer the same problems of comparison, but do allow at least some control over extraneous factors. Communications Division looked at " before" and "after" releases from three regular economic series: building permits, capacity utilization and provincial economic accounts.

Before redevelopment, the building permits release attracted one minor news service story in one newspaper among those monitored by Statistics Canada. After a major overhaul, the release received substantive coverage in the two leading business newspapers and was cited in a broader article by the economics columnist for the leading newspaper chain. Moreover, there were substantially more direct quotes from the more recent release than there had been before.

Like building permits, the release of industrial capacity utilization data had previously attracted only one story in the leading financial newspaper. After redevelopment, it received substantive coverage in eight daily newspapers, including Page One coverage in the widely read business section of the daily that claims to be "Canada's national newspaper."

The provincial economic accounts release was an exceptional case. Before redevelopment, it had attracted two stories in major daily newspapers, partly because of confusion with a related release. In redevelopment, the two releases were combined and, because of its regional interest, a special effort was made to bring the release to the attention of regional media. The effort paid off in 22 articles in major daily newspapers and 15 newscasts among the media outlets monitored by Statistics Canada.

Without being conclusive, these three case studies suggest that, other things being equal, the new style of release has in fact increased both the number and length of articles. In terms of quality of release coverage, the analytical information provided in *The Daily* is clearly, if selectively, being carried forward in media reports, often as direct quotes.

4.4 Additional Benefits

A more positive, even enthusiastic, reception for journalists contacting analysts has been an added benefit. Increased contact with working journalists in unthreatening circumstances has made analysts more willing to provide additional clarification or commentary on a release. This greater collaboration between analysts and journalists has also expanded and improved the coverage of releases, as well as increased the use of Statistics Canada's information as background to other stories.

5. CONCLUSION

The changes introduced in *The Daily* appear to have been successful in inducing the news media to convey more Agency information to the Canadian public in a form that demonstrates its relevance and significance. Implicitly, Statistics Canada has succeeded in using the news media to better inform the Canadian public.

Our contacts with other national statistical agencies suggest that a number of them are also attempting to make similar improvements to their releases. The Statistics Canada model may provide, or at least suggest, the means to bring about radical change. It was not the result of a detailed, *a priori* plan. Rather, it evolved by trial and error. Things that worked were built on, things that did not were abandoned. In the end, it was clear that neither a top-down nor bottom-up approach alone would have achieved the desired results.

At another level, the Statistics Canada initiative provides a fascinating case study of the process necessary to achieve radical change quickly in a decentralized environment. It illustrates the multiple levels of support and the sustained energy necessary to alter corporate culture. Many of the lessons are applicable to quite different projects. Knowing what needs to be done is not always enough to make it happen.

SESSION 7

Data Warehousing

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

NEW TECHNOLOGIES FOR DATA COLLECTION AND DISSEMINATION

W.J. Keller and W.F.H. Ypma[1]

ABSTRACT

This paper gives a brief description of some of the information-technological developments within Statistics Netherlands. After an overview of the effects on the production process it focuses on two aspects, Electronic Data Interchange (EDI) for datacollection and the Internet for dissemination. Among the many projects currently running at Statistics Netherlands "EDI Pilot 2" is described. This concerns EDI on the financial accounts of enterprises. We will also discuss our so-called dynamic Web page system called WITCH. Both for data collection and dissemination, we will focus on the role of the meta-information as a tool to control the process. We will see how technology changes this role and generates new possibilities to enhance the effectiveness of the meta-information.

KEY WORDS: Official Statistics; Datacollection; Dissemination; EDI; Internet; Meta-information.

1. INTRODUCTION

Statistics Netherlands is at present under the influence of several developments. As everywhere else it no longer operates as an untouchable organisation of civil servants. Efficiency and market-orientation are the key-words now. We need to produce at lower costs. Furthermore we need to lower the costs we inflict upon our suppliers of data. The outcome should be a product that, although not actually sold on a market, our clients eventually want.

Furthermore we are confronted with new developments in Information Technology (IT). They will give us the opportunities to construct the necessary tools to meet the new demands. In a situation like this a NSI needs to make the right strategic choices.

2. DEMAND-PULL

The production process is on the one hand influenced by the growing demands of our clients and respondents. There is a strong political demand for a decrease in the respondent burden as a part of alleviating the administrative burden of enterprises. Statistics Netherlands sends out 1.25 million questionnaires to enterprises and other institutions per annum. Large and medium-sized enterprises may receive as many as 50 questionnaires per year, including repetitive monthly and quarterly surveys. In particular larger companies in manufacturing are subjected to many (about 20) different types of surveys. The conclusion is clear: Statistics Netherlands has "to fight the form-filling burden". Furthermore, budgets are shrinking so there is a demand for higher efficiency and higher productivity.

Concerning our output we see a demand for a higher user-friendliness. One particular aspect is a demand for an improvement of the coherence of the totality of the information we offer. Another aspect is that our clients will want to be able to use the new media IT has to offer.

3. TECHNOLOGY PUSH

On the other hand we are blessed with information-technological (IT) developments or the technology push. In the first place these developments give us new technical possibilities, the means to construct new tools for our production process. We see large improvements in the possibilities of data processing, data storage and

[1] W.J. Keller, Statistics Netherlands, Director of Division Research and Development, P.O. Box 4000, 2270 JM Voorburg, the Netherlands; W.F.H. Ypma, Department Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, Netherlands.

data transmission. The last aspect will probably have the most striking influence on our work: the communication of data between our respondents and the NSI on the one hand and the communication of data between the NSI and its clients on the other.

In the second place these new developments create their own demand. The new technology will be used anywhere. Our suppliers of data will use it. Our clients will use it. They will no longer be satisfied to communicate with us in the old way, that is on paper. Our suppliers produce their data by electronic means and will want to use those means to deliver those data directly to us in order to minimise their own costs. Our clients process our data by electronic means. They will demand to be able to select and receive those data with the tools that IT has to offer.

These two factors lead to the conclusion that the NSI will have to make those strategic choices in its production process that make the best use of the possibilities IT has to offer.

4. STRATEGIC CHOICES

New demands and new tools will affect all the aspects of our production process. To describe them let us first discern, within this production process, three stages. The input-phase is where the data are collected in contact with the respondents. In the throughput-phase these data are processed to produce the information with the characteristics we are actually looking for. In the output-phase this information is offered to and disseminated among our clients.

Let us begin with the input-side, the collecting of data. First, data-collection among individuals and households. It is not saying too much when we state that a major step forward has already been taken at Statistics Netherlands. We have introduced all kinds of Computer Aided Interviewing (CAI) and developed BLAISE to do so. (Needless to say that BLAISE does more than develop and present electronic questionnaires.) The gains of these developments was mainly in terms of an increase in productivity or efficiency. The number of staff needed for coding, data entry and checking decreased dramatically. This efficiency also shows itself in the much faster production of results. Still, there is even more to gain. In the first place on the efficiency of the production process itself. But also in the statistical sphere improvements are still possible: new ways of interviewing: CASI, computer aided self interviewing, and, not directly a matter of IT, more efficient sample designs.

Much more however is still to be done in the field of collecting data among enterprises. The demands here are stronger. Response burden has become an issue. It is the driving factor behind our strategic choices here. When we see at the same time that almost everywhere automation and IT has invaded the bookkeeping systems of the respondents involved, it is clear what our task for the nearby future will be: the Edi-fication of the collection of information from enterprises by the NSI. What CAI is for interviewing among households, EDI (electronic data interchange) will be for datacollection among enterprises. Later in this paper we will go deeper into EDI with enterprises.

In the throughput phase we are looking for more efficient ways of processing our data. Of course CAI and EDI make much of the editing superfluous. Less errors will be made. Still we expect much from more efficient or rational ways to handle the editing process. Here data processing is the key. The choice will be that we will no longer edit each individual record. It should be possible to use the computer to find the worst errors and help to correct them. At the same time the computer can prevent us to spend time and money on correcting unimportant errors. The gains here are in the first place productivity gains.

Finally, the output phase. Here the new developments probably get the most attention from the public. We see the new media by which information can be presented to its users. Paper publications may continue to play their role but especially the more professional user will want to select and receive his data by electronic means. Statistics Netherlands is producing or developing those means: data on CD-ROM, data on Internet. More important and maybe more difficult is the way data should be presented with those new media. The amount of information will be much larger than we had in our paper publications. At that point the management of the meta-information becomes crucial.

For this purpose Statistics Netherlands is developing STATLINE. This should lead to a data-base intended for the end-users that should give access to "all" our data. As could be expected, structuring those data is the main problem. At the same time we are confronted with lacking coherence due to lacking statistical co-ordination. STATLINE is intended to play a key-role in the dissemination process of our data. The strategic choice has been made that we aim for that structure wherein all publications and all other dissemination of data goes through STATLINE.

In section 11 and later we will discuss STATLINE in more detail, in particular in relation to the Internet.

5. RESTRUCTURING THE PRODUCTION PROCESS

In the previous section we described the strategic choices we made regarding the different phases of our production process. Those choices go further than just the development of a new tool. They will affect the structure of the production process itself. One should be prepared to take those consequences as well. The present or the "old" way the production process is structured is along the lines of the individual statistics. For each statistic - an end-product - a new questionnaire is designed, respondents are selected, data are processed and a publication is made. Especially on the input side this is inefficient.

In the new situation, we are talking more than 10 years from now, especially the datacollection will be re-ordered. No longer the demand for information but the supply, the available actual data-sets, will dictate the organisation there: the sources. Each source will be tapped once and completely for any possible use within the NSI. The collection is technically and conceptually adapted to that source. (In the remaining sections of this paper we will give some indication regarding the nature of those sources.)

Having collected the data we may have to translate them to statistically suitable concepts, integrate them and we will have to distribute them among users. They may be inside the NSI, the integrative systems like the National Accounts, or outside the NSI. This means that somewhere those data will have to come together for distribution. For the input-side this can be illustrated as follows:

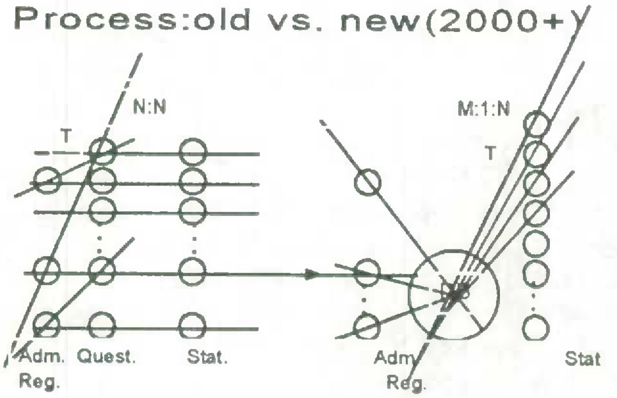


Figure 1

On the left we see the old situation with a separate production line for each individual statistic. On the right the future situation. There, all the possible sources contribute to a central database of relevant information. From that database the actual statistics are produced by combining the relevant information. It is evident that in order to combine information one should be certain that the characteristics of that information are such that combination makes sense. Those characteristics are specified in the meta-information.

6. ELECTRONIC DATA INTERCHANGE (EDI)

From now on we will focus on EDI with enterprises and institutions. A NSI collects data to produce statistical output. What needs to be done is making a translation from the data of the respondent to the data of the output. This is done in several steps. The first step may be left to the respondent. If so, it leads to a certain response burden.

The first step of the translation involves two parts. First there is the conceptual translation, the mapping of the concepts of the source, the administrative concepts, on the concepts to be delivered to the NSI. This is the most difficult part. Not only do business records differ from statistical information but also do they differ among themselves. The second part of the translation is a technical one. We would like to receive data in a suitable technical form. Especially we and our respondents would like to avoid data-entry.

7. MODES OF EDI

Electronic data interchange will be one of the strategic tools to meet the challenge of lowering the response burden and improving our productivity. In every individual case we should decide whether to use it and in what mode. We will describe several modes of EDI and judge them by their effect upon the response burden. Of each possibility we will indicate the nature of the translation and especially who is going to make it. We concentrate on the conceptual translation.

7.1 EDI on centrally kept registers

Here we do not approach the individual respondent at all. We are dealing with centrally kept information on individual units, collected for other purposes than statistics and yet of interest to the statistician. In itself this way of data collection creates no response burden.

There are however disadvantages. The most important is that there is very limited choice as to the conceptual contents of the data the NSI receives. In other words one cannot ask for much translation towards statistical concepts. That will have to be done by the NSI itself.

The second problem is closely connected and is that of units and populations. Here also one cannot but accept what the register keeper is able to supply. If the units he uses do not comply with the statistical units there is a problem. The same is true regarding the classification of those units. How can we connect the register population to our total statistical population?

A third problem regards the sampling strategy. If the register provides us with yearly data on, let us say, 70% of a population we formerly used to describe with a rotating sample of 1 out of 5, then what should our strategy be regarding the remaining 30%?

In the Netherlands there are several examples of usable registers. There are centrally kept registers of enterprises with the chambers of commerce. The tape of these registers feed our own register of statistical units. Statistical data can also be had from fiscal (company tax, VAT) or social security sources. For several possibilities (chambers of commerce, company tax and VAT) the possibilities are used or being researched.

7.2 Commercial bookkeeping bureau's

A related possibility is tapping from the information of commercial bookkeeping bureau's. They keep the records on financial information or regarding the wages of sometimes a large number of individual enterprises. This possibility also is attractive because of the large number of respondents involved with only one link. Furthermore these service bureau's will be capable of providing us with more information than e.g. the fiscal records contain. A disadvantage is that these service bureau probably will charge their clients for answering the questions of the NSI. Not every client will be prepared to pay.

Having said that these bureau's often hold much of the information the NSI needs, there are two possibilities regarding the question who will make the translation. The answer is a matter of cost benefit analysis. There is an example at Statistics Netherlands of one bureau that does the bookkeeping of 40% of the enterprises in one particular branch. In that case it is profitable for the NSI to make the necessary translation. In other cases we propose to provide software by which the bureau itself makes the necessary translation.

7.3 EDI on individual respondents

When the above described possibilities are not available we will have to approach the individual respondent. In doing so we should be aware of the fact that sometimes we will have to discern within one statistical unit, often an enterprise, several sets of administrative records. We will see that we will have to approach these subsets separately and in a different manner. Within commercial enterprises we find the financial records, the logistical information (foreign trade, stocks) and the records on wages and employment. Especially the financial records and those on wages are strictly separated in the Dutch situation.

Here we classify by the translator of the information.

7.3.1 The NSI translates

One of our EDI-projects - EFLO - works along this line. It deals with the data from the Dutch municipalities. They deliver a set of records directly tapped from their own complete set of records. The translation is done at Statistics Netherlands. The advantages in terms of respondents' burden are evident. Although extra work by the NSI is needed, this extra work can be seen as an investment depending on the stability of the translation scheme. It is expected that this form of EDI will lead to an improvement of productivity once the translation schemes are completed. Important is that we are here dealing with a limited number (600) of respondents.

7.3.2 The respondent translates to a standard record

Here a standard record of information is defined. The standardisation regards both the conceptual and the technical aspects. To produce the record, writing the software, is left to the respondent. Working with a standard record is not always possible. It can only be done when the information is already standardised among respondents to a certain degree. Furthermore, to make a standard record possible the NSI sometimes may have to move towards the concepts of the respondent. In that case a larger part of the total translation to the final statistical output has to be done by the NSI.

Especially when the standard record is available in the bookkeeping software the respondent uses and regularly updates, this mode of EDI has a clearly favourable effect on the respondents' burden.

There are two examples. One is IRIS, the EDI on intra-EC trade. The standard record developed here is implemented in over 40 software systems available on the Dutch market, after certification by Statistics Netherlands. The EGUSES project is the other example.

It regards wage information. That subset of company records is highly regulated in the Netherlands. That fact made it possible to define a standard record.

7.3.3 The respondent translates. No standard record

Still a very large part of the information we are looking for is left out. The respondent has it in a form that conceptually and technically differs from what the NSI wants and from what other respondents have. We distinguish:

- Paper questionnaires

This clearly is no form of EDI. We mention it as a possibility to be complete and to emphasise the point that here the respondent does all the translating by himself and each time has to do it all over again.

- Electronic questionnaires

Although strictly speaking at most partial EDI, this method proves very successful with IRIS, the software on INTRA-EC trade. (IRIS works with a standard record as well as with data entry.) By providing extra help-functions and the possibilities of adapting the questionnaire to the individual respondent it also helps to lower the response burden.

- "Full" EDI

The last possibility is that the NSI provides the software by which the respondent can set up a translation scheme for both the technical and the conceptual translation. Once set up, and in so far as no changes occur, the scheme can be used to produce data to be delivered to the NSI. The example here is EDI-Pilot 2 directed at the financial records and described in the next section.

Before we go into that, we give a summary of the characteristics of the several possibilities of EDI on individual enterprises:

(Sub)sets of records	Financial Wages Logistics All records
Translator	NSI Respondent
Output of Respondent	Not translated data Standard record Non-standard record Data entry: electronic questionnaire paper questionnaire

8. EDI-PILOT 2

We will now describe the project EDI-Pilot 2 directed at the financial records of individual enterprise as an example. It shows the problems one has to face. While describing Pilot 2 we can refer to the scheme in the previous section.

Pilot 2 is directed towards individual financial accounts. In the Dutch situation these are only a part of the accounts of an enterprise. Especially the accounts on wages and employment are excluded. This is not a choice voluntarily made by Statistics Netherlands but one forced upon us by the way the bookkeeping systems are organised in our country. Leaving out detailed questions on wages, we combine within Pilot 2 all the questions that are put to the financial accounts. The result is the combined questionnaire.

The contents of the combined questionnaire are dictated by what is available in the financial accounts. Regulated as our society may be, the financial accounts may diverge strongly in internal organisation and in the concepts used. In the first place this means that we will have to adapt our questions towards the possibilities of the automated system of the enterprises. This may imply more statistical work for the NSI to reach the same output. If one wants more, it will probably be necessary to ask for additional information to be given explicitly by the respondent, that means by data-entry. In the second place the diversity of respondents means that a unique translation scheme will have to be set up and maintained for each respondent.

Financial accounts also differ in their technical lay-out. A large number of bookkeeping software systems is in use. There is no standard record for information to be selected electronically from the software and it is not expected that it will be possible to define one within the near future. As the main goal of Pilot 2 was the lessening of the respondents' burden, it was decided that the amount of data-entry was to be minimised.

That means that some ingenuity was needed to create the automated link we were looking for. This is done by using the reports or print-outs of the software system. Instead of printing them, they are sent to a file, a print-file, to be read by the translator, the main part of the software module that will run on the respondents computer that is now being developed as part of Pilot 2. The layout of the reports and thus of the printfiles is fairly stable. The respondent communicates this lay-out to the translator. He defines rows and columns within the report. Subsequently he tells the translator how to manipulate the rows and columns in order to transform the information in the report to the statistical information asked for by the combined questionnaire. The resulting records are sent over to Statistics Netherlands.

The Translator

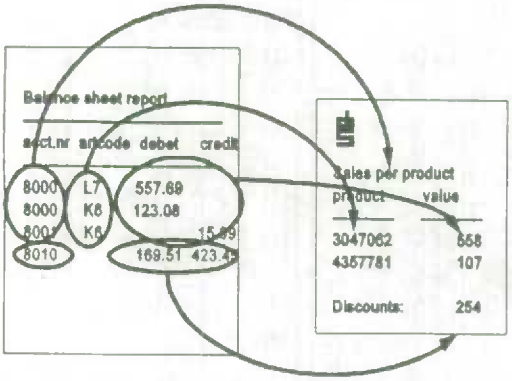


Figure 2

We see then the two parts of the translation scheme. The first part lays down the lay-out of the printfiles to make the technical transformation. The second part defines the conceptual transformation of the information to be found on the printfile towards the statistical information asked for on the combined questionnaire.

The final question is who will make that translation scheme. One of the principles of Pilot 2 is that "the respondent translates". This means that the respondent himself has to set up the translation scheme. This of course make it less respondent friendly. It seemed however impossible to set up those translation schemes at Statistics Netherlands. It is clear that this is not an easy task for the respondent. On the one hand this means that a strong help-desk and a fairly large field service is needed, and on the other hand this means that even with Pilot 2 we will not yet reach the ultimate user-friendliness of EDI.

We expect the translation scheme to be fairly stable or, in other words, that technical and conceptual changes will not be too frequent. A second time the translator can use the already available translation scheme to produce the statistical information. Answering the combined questionnaire then becomes a matter of minutes instead of hours and can be handled by a less qualified employee. That is what makes the concept attractive and the initial investment worthwhile to the respondent.

9. SCOPE OF PILOT 2

As said, Pilot 2 is directed towards the financial accounts. The principle is that all the information that is tapped from the financial accounts by any statistic of Statistics Netherlands will go through Pilot 2 if automated retrieval of that information is possible. In practice this means that several large statistics will switch completely to EDI. For industry, our main target, we find:

-Monthly statistics on total turnover
-Monthly statistics on foreign trade, by product
-Quarterly statistics on turnover by product
-Yearly statistics on gross investment
-Yearly statistics on the production process
-Yearly statistics on the financial processes, inc. balance sheets

The participation of foreign trade is a pilot within the pilot. Not only does Statistics Netherlands already have a successful EDI on this area in IRIS, but also the possibilities of getting enough foreign trade data when aiming in the first place at the financial accounts, still have to be researched.

Some questions in the above mentioned statistics are dropped, e.g. the questions on quantities of energy used in the production statistics. They cannot be addressed by this form of EDI. Probably a separate paper questionnaire on this subject will be sent.

On the other hand, some questions originating from other statistics mainly aimed at other subjects and accounts (e.g. the labour and wage accounts) are included because the answers are typically to be found within the financial accounts of the enterprise.

The domain of EDI consists of those commercial enterprises that have set up financial accounts by means of computer software that satisfies certain technical specifications. In practice this means that we direct ourselves towards the profit sector within industry, trade and services. We start with industry because there the gains in terms of lessening the respondents' burden will be the largest. Individual smaller enterprises are not included because their bookkeeping and automation capacities are expected to be too low. In view of the relative small amount of information asked here, more is expected form centrally kept records (VAT, corporate tax) and from bookkeeping bureau's often keeping books for hundreds of smaller enterprises. The very large enterprises are also excluded. Because of there complexity they need an individual approach of course in the end also by means of EDI but then "tailor made".

Regarding the number of respondent participating in this kind of EDI, we should mention that in pilot 1 a number of 12 respondents participated and still do. Pilot 2 will start with a field test next march aimed at 20 respondents. Starting September 1996 we aim at larger numbers. By end 1996 Pilot 2 should handle several hundreds of respondents. Pilot 2 will also be used to approach the bookkeeping bureau's. That will lead to larger numbers of statistical units described with one EDI-link. If EDI-Pilot 2 is successful we will, following pilot 2, in 1997 aim at a number of 25,000 units to be approached with this instrument, partly through the bookkeeping bureau's.

The revenue of Pilot 2, if successful, will in the first place be a relief of the respondents' burden. Productivity gains will not be that large. In the first place all kinds of activities remain. Not every respondent will participate, data will still have to be checked etc. In the second place new activities arise in the form of a growing help-desk and a field-service that will not only have to cope with bookkeeping problems but also with technical automation problems.

10. CONTROLLING PILOT 2: THE META-SYSTEM

Eventually Statistics Netherlands aims to reach several thousands of respondents. This of course asks for a control system to deal with the production of the appropriate questionnaire, sending it to the respondent, checking the, timely?, response, checking and storing the incoming data and controlling possible feed back etc. This means that a lot of information, meta-information, on the respondents has to be kept updated.

Another part of the meta-information deals with the contents of the combined questionnaire. As an example we will focus on that part.

Constructing the combined questionnaire needs to co-ordinate the approach of the different statistics aimed at the financial records among each other but also with the bookkeeping practices of the respondents. Of course the latter already happened before but with EDI it will become more explicit. This needed some negotiation. It is clear that with EDI up and running, much of the former autonomy of the individual statistics, especially regarding their questionnaire, disappears.

The module containing the translator gives us better opportunities for supplying meta-information to the respondent than before. There are the usual on-line help-functions. By means of hypertext the explanations are linked. For the help-desk and for the field service probably a more detailed system of help-functions and explanations will be set up. The system not only contains cross-linkages but also simple computational rules so that for instance totals can be computed.

For this end a set of variables was laid down in a database, with names, questions texts, explanations and, if necessary, computational relations with other variables. From this database, variables, question-texts, explanations etc. are selected and combined to questionnaires. Respondents are classified into clusters by size, branch of activity and type of financial records kept. Sometimes sale-records are kept by the enterprise itself but the yearly balance sheets are set up by a bookkeeping bureau. For that statistical unit the total of the information needed will have to be collected by two different questionnaires directed towards two different reporting units. Each cluster gets its own combined questionnaire.

11. EDI: CONCLUSIONS

In this way a large set of meta-information on concepts emerges. This meta-information controls the process of datacollection. A question aimed at the financial records can only get there through the central database of variables. When entering the variable, the relation with the rest of the contents will have to be made clear. It has to fit in.

In the first place we now see that the character of meta-information has changed. In most of the literature we often find meta-information as a mere descriptive piece of information only available if the statistician has found the time to set it up, mostly after he has produced his statistic, for the benefit of the user. If later on the statistician diverges from his earlier meta-information there is nothing to stop him and nothing that guarantees that the meta-information will be adapted.

Here we find a piece of meta-information that has to be set up before the production process starts. The statistician cannot but use the meta-information system. The meta-information has become a tool in the production process. From being descriptive it has come to be prescriptive. Earlier we saw the same thing happening with datacollection among households through BLAISE.

This however has further reaching consequences. We can now go back to the first sections of this paper. There we spoke of the extra demands put to Statistics Netherlands. One of them was less respondents' burden.

That was the first goal of EDI-Pilot 2. But we also see here how the technology push gives us some opportunities to answer another demand namely that for more coherence. It goes without doubt that the way EDI is implemented here will lead to a larger extent of statistical (conceptual) co-ordination. We mentioned the power of the meta-system and we also see that within EDI a number of statistics is combined that were earlier produced in separate, independent processes. Remarkable is the fact that this growth in statistical co-ordination is not reached by an increase in central directives but as a side-product of the tools used in the production process. We do not think that all the problems of the coherence of our end-product, that means all the problems of statistical co-ordination, can be solved by devising the proper tool. We do think however that further improvements can be made in this field by applying the possibilities of the technology push in the right way.

12. DISSEMINATION

Let us now turn to the output side of the statistical process. At present, most statistical agencies provide aggregated statistical information in various ways, but dominantly in printed form (on paper). Because printing is a relative cumbersome and expensive way of dissemination, more and more people are looking at the electronic highway (a.k.a. the Internet) as a cheap and easy way to disseminate statistical information. This paper focuses on the impact this trend has on official statistics. We will argue that besides the technological dimension of publishing on the Internet, the main problems will be conceptual, i.e. those of statistical coordination and integration.

In this paper we will discuss some projects at Statistics Netherlands (SN) dealing with electronic dissemination. We will cover Statline (our statistical database with a traditional on-line query tool on a remote DOS client) and its new experimental version, Statline-WITCH, with so-called dynamic Web pages on Internet. We will argue that by combining the ease of use and ease of access of the Internet with the multi-dimensional database systems found in statistics, great opportunities for statistical dissemination will arrive.

13. ELECTRONIC PUBLICATION AT THE INTERNET

Presently, our publications take many different shapes: printed paper, floppy disks, faxes and CD-ROM's, automatic and human voice response, press release, videotex, etc. Behind all these different media there is (aggregated) statistical information, often in machine readable form e.g. as the output of survey processing systems. Needed is a "one-stop" dissemination database situated between the internal processing systems and the outside world, capable of producing many different media from one source, in a consistent, timely and efficient way. Besides on-line access to the database, such a database system could also automatically provide the information for other media, such as floppies, E-mail subscriptions, faxes and CD ROM publications. But one of the most important objective of such a system is to provide easier on-line access by our customers to the wealth of information at statistical bureau's. It is our opinion that in this respect the Internet will play a very important role in the near future.

With the rise of the graphical browsers (Mosaic, Netscape) on the Internet, the net has grown immensely during the last year. Within months, nearly every respectable company has set up its own so-called "Web-server" on the World Wide Web (WWW). The net, with its sky-rocketing popularity and therefore great infrastructure, is already connecting tens of millions people all over the world, with access becoming easier and bandwidth nearly free (in Holland, the Internet, at 28 kbps speed, will be a local phone call away for nearly everyone at the end of 1995). It allows statisticians not only to collect information more efficiently (see our paper on EDI), but also to disseminate aggregate statistics more efficiently, with a marginal reproduction and distribution price close to zero. (There are already 27 000 free Internet subscribers to David Letterman Top Ten List server: imagine such a circulation to our press releases !)

At present, several statistical institutes publish information on the net through the WWW. Well-known Web-servers are those from the US Census Bureau, Statistics Canada, Eurostat and SN, to name a few. Everyone with an Internet connection and a browser like Netscape can visit these servers from all over the world. Most of the material published on these Web's, however, is not really statistical information, but lists of publications, press releases, and general information for the public. The limited amount of truly statistical figures is often presented in a documentary way, i.e. as electronic copies of the printed pages from traditional publications.

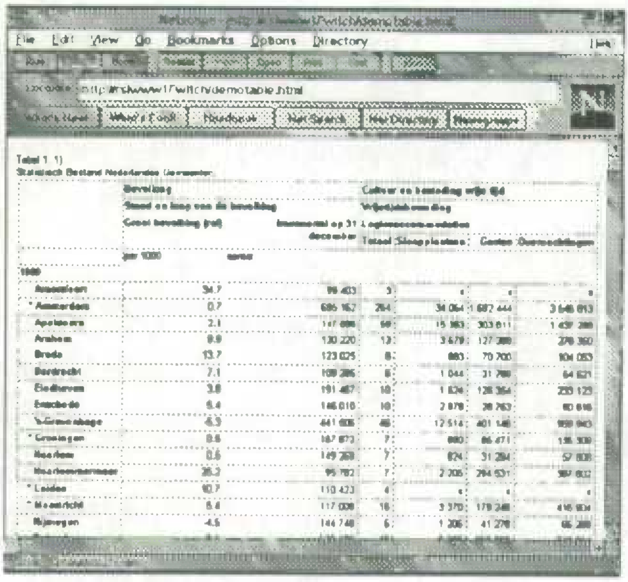
This approach, which is typical for so-called *static Web pages*, makes it difficult to manipulate statistical figures as structured information, since the user only has access to documents, i.e. (formatted) text. What is really needed is access (through the Internet) to a real *database*, encompassing various statistical sources in an integrated system. Once our statistical information is available in a structured, machine readable way, we can manipulate it and present it in any form, including unstructured (like a text document) and structured (e.g. like a spreadsheet). This structured database approach is also necessary in order to be able to provide better coordinated and integrated statistical information.

An example of a statistical database is Statline, from SN. Statline is based on the client/server concept, where the front end (running on a PC, possibly outside SN) is separated from the back-end (the database server, located at SN). Front end and back-end are presently connected through traditional datacommunication facilities like Local Area Networks (LAN's) internally or simple asynchronous lines (using telephone lines and modems) externally. In order to optimize the performance of its multi-dimensional database (see section 4), Statline uses a proprietary, non-relational database design based upon indexed files. The DOS-based front-end uses a user-friendly window/mouse desktop metaphor where the results of searches are displayed in a type of multi-dimensional spreadsheet, with additional graphical views, including thematic maps. The Statline front-end is the same as the software we use as interface to our floppy disk (or CD-ROM) - based publications. Presently, Statline does not use the Internet, but this will change when we introduce the concept of *dynamic Web pages*.

14. DYNAMIC WEB PAGES: COMBINING INTERNET WITH DATABASES

As discussed above, the statistical information found on ordinary Web pages on the Internet is difficult to manipulate in a structured way, in view of the documentary (non-numerical) character of a Web page. Also, each Web page is static in nature, i.e. we have to prepare each page beforehand by storing its (documentary) image on the Web server. Wouldn't it be great to combine the power of on-line databases, like our Statline database, with the ease of use and access of the World Wide Web? This is where the so-called *dynamic Web page* enters the picture. The idea is to use browsers like Netscape as front end to systems like Statline. Each time a user requests data, a special interface, called WITCH, translates the request to the Statline format and generates a Web page on-the-fly to present the result from Statline to the user.

An example of a WITCH generated Web page, using Netscape 1.1 with HTML3 table-support, is shown below.



Tabel 1.1)
Statistisch Bestand Nederlandse Gemeenten

	Bevolking		Cultuur en besteding vrije tijd			
	Stand en loop van de bevolking		Vrijetijdsbesteding			
	Groei bevolking (rel)	Inwonertal op 31 december	Logiesaccommodaties			
			Totaal	Slaapplaatsen	Gasten	Overnachtingen
	per 1000	aantal				
[illegible]						
Amersfoort	34.7	99 403	3	.	.	.
* Amsterdam	0.7	686 162	264	34 064	1 687 444	3 646 013
Apeldoorn	2.1	147 [illegible]	69	15 363	303 611	1 437 [illegible]
Arnhem	9.9	130 220	13	3 679	127 [illegible]	270 360
Breda	13.7	123 025	8	883	70 700	104 083
Dordrecht	7.1	109 285	6	1 044	31 780	64 621
Eindhoven	3.8	191 467	10	1 624	128 364	230 123
Enschede	5.4	146 010	10	2 878	38 763	[illegible]
's-Gravenhage	4.3	441 [illegible]	46	12 514	401 [illegible]	[illegible]
* Groningen	0.6	167 873	7	880	86 471	135 [illegible]
Haarlem	0.6	149 269	7	824	31 284	57 808
Haarlemmermeer	26.2	95 782	7	2 206	294 531	397 832
* Leiden	10.7	110 423	4	.	.	.
* Maastricht	5.4	117 008	16	3 370	179 248	416 [illegible]
Nijmegen	-4.5	144 748	6	1 206	41 278	66 288

Figure 3

By using a Web-browser as front-end to a database with structured information, other Web-tools also become available. For example, besides presentation in a Web format, we can also download information or use other "viewers" in the browser, e.g. to see spreadsheets, graphs, or maps from the net. WITCH will not only generate dynamic Web pages but other formats like spreadsheets as well. In this way, the user can save information in a structured format in order to manipulate the data later.

The advantages of this approach are several: first, we don't have to build our own front end tool, like we did with Statline for DOS. Anyone with a decent Web-browser can access Statline, wherever in the world. Second, by using the commonly available Web browser, Statline becomes immediately available on different platforms (Windows, Mac, UNIX). Third, the user does not has to learn a new interface, once the Web browser is known. Finally, we can use the Internet as communication medium, with all its advantages: high bandwidth (28.8 Kbps by modem or even better in case of ISDN or T1 links) and great accessibility (as said before, in the Netherlands the Internet will be a local phone call away for nearly everyone at the end of 1995).

With millions of potential users being able to access our on-line statistical databases over the Internet, new challenges arrive. The biggest concern, as we see it, is the statistical coordination of the information we provide.

15. STATISTICAL COORDINATION

Most statistical bureau's provides hundreds of different statistical publications from several hundreds of surveys. All this amounts to million of figures, thousands of tabulations, and many, many different sources of information. But except for some special publications (like the National Accounts), each publication only deals with a very specific topic, and users are confronted with an inaccessible "gold mine of information" with many, many different faces. Someone being interested in, say, automobiles, has to look in more than a dozen publications to get a total picture, encompassing the production of cars, the exports and imports, the use (in time and mileage), the energy consumption, traffic accidents, the environmental effects, etc. Finding all this information can be laborious and troublesome, especially since each statistical department focuses only on their topics and publications. At the same time, NSI's sell only a very limited number of copies of each individual publication, often without recovering the full dissemination costs, let alone the collection costs. And finally, while users appreciate our impartiality and accuracy, they complain about the lack of timeliness of our statistical information.

If all available statistical information is placed on the Internet, free of charge, millions of users can and will access it. Compare this with the hundreds of users reading our printed publications. However, not only the implications in terms of distribution are mind dazzling, also the conceptual implications will be great and probably very problematic. Why? Since with such a unlimited access to all statistical information, users will ask much better access paths (with search by keyword and multi-dimensional queries, on time, branch, region, etc.. on top). And then, after we have provided these tools, they will find out that our information is not always coordinated, let alone integrated. Inconsistencies, buried in hundreds of different paper publications, will become visible on the net, and users will start asking questions: not only for more, but also for better coordinated and better structured information.

One answer to this demand for better coordinated data is the systems approach, like National Accounts. Another, less ambitious goal, is to coordinate the classifications, domains and definitions used in different statistical publications. This is the philosophy behind a new database approach, based on the concept of multi-dimensional tables or *cubicles*.

As in Computer Assisted Interviewing (CAI) systems, we can distinguish between the data itself and its description, the so-called metadata. While the CAI systems focus on the individual data and metadata processed in the data collection and editing stages, in the dissemination database we focus on the aggregated data and its metadata. This metadata comprises both the syntax (format) and semantics of the data (the definitions of the published variables, like the definition of "number of employees"), as a description of the survey itself, the sources and how the items are derived. The first step to coordinate statistical publications is to standardize the definitions of the variables used.

Each item (e.g. number of employees) is often available for different domains, defined by crossings of discrete, categorical variables, like sector, region or time. An other important mechanism to coordinate the dissemination of statistical data is the standardization of these categorical variables, leading to classifications e.g. for branches of industries, commodities, regions, etc. The basic representation of information used in such a database is therefore the multi-dimensional matrix (sometimes called "cubicle") where one dimension reflects the different variables (e.g. number of employees, profit, prices), a second one the (discrete) time axis (e.g. years and months), while other dimensions correspond to various classifications (industries, commodities, regions, etc.). The items inside the matrix reflect the measurements ("number of") on a certain variable ("employees") in the domain defined by the crossing of the categories on the other axis ("in industry x in region y at time t"). Often, categories are classified into different systems of detail (e.g. a n-digit industries classification, with n=1..9) which are often (but not always!) hierarchical to one another, resulting into levels of classification.

Metadata (descriptions) in this database of cubicles can refer to the total matrix, to the axes and their variables and categories, and to the individual items inside the matrix. Particular problems of metadata arise when the definitions of certain categories (like regions, industries, commodities) change over domains, in particular over time. As example, take a region like a municipality. Not only the number of inhabitants in Amsterdam in 1980 is different from 1990, but also the definition of Amsterdam itself differs between the two years (e.g. because of border corrections)! Similar problems arise when certain items are only available for certain categories or classification levels, making comparison of different items in various domains sometimes impossible.

As explained above, in statistical databases the most important type is the multi-dimensional object, or *cubicle*. A statistical database will contain many, many different cubicles, which might all share similar classifications along some of their axis. Besides these multi-dimensional objects, also simple ("flat") two-dimensional cross-tabulations, as shown in most traditional statistical publications, have to be stored and presented in the database, as well as (one dimensional) text objects like press releases. All this information is documented (metadata) inside the database on various levels (from the total object down to the individual items or cells). A classification of database objects into the well-known statistical domains (like economic, social and demographic statistics), and classifications thereof (production, environment, labour-market, well-being, etc.), make navigating through this immense database of information more feasible. A very strong tool in finding the information needed is a database-wide keyword (thesaurus) system, which allows the user to quickly allocate the right object.

In several countries, statistical databases based on this concept of cubicles are presently being used or under construction. Well-known systems include PC-Axis from Statistics Sweden, the ABS-Database from the Australian Bureau of Statistics and the above mentioned Statline database from SN. Statline is both an internal SN system as well as an open on-line database system, available for customers outside SN. Inside SN, it will be used to "drive" the publication of different media (paper, floppy, videotext, etc.) as much as possible without any human intervention (e.g. with fully automatic composition of printed material). Stateline is also used for all our internal inquiries, e.g. for customer support, and to provide customers with snapshots from Statline on a incidental or regular basis (e.g. with automatic fax/E-mail subscription on "hot" figures). It will also, and possibly most important, be used as a vehicle to standardize (and therefore coordinate) all our aggregated statistical information, including our metadata.

Outside SN, Statline provides a direct connection for our large accounts to the wealth of information Statistics Netherlands provide. By combining the WITCH interface, using the idea of dynamic Web pages, we will allow for easy access to Statline over the Internet to many other customers.

16. DISSEMINATION: CONCLUSIONS

With the spectacular rise in the use of the Internet world-wide, electronic publishing quickly becomes a reality. The Internet and in particular the WWW (World Wide Web) not only provides great ease of use and ease of access to an immense universe of information, it also provides great challenges to statisticians. Should we simply put all our paper-based publications in electronic form on a Web server, using the same document form as we did in printing? Or should we make our statistical information available in a more structured way? We think that the technology of the so-called dynamic Web page as a front-end to a database with statistical figures will be a better solution than static Web pages, which in some way just replicates the paper metaphor.

More in general, once statistical information is available in a structured, machine-readable format like Statline, we can present it in any form by just using interfaces like WITCH. From such a database, not only Web pages can be generated on the fly, but also fax/E-mail messages, press releases, databases on CD-ROM and even old-fashioned printed output in a completely automated way. Of course, not only the data itself but also the metadata should be machine readable, including the syntax (format) and the semantics (content) of the data. Once this is achieved, we can easily exchange information between statisticians using standard export formats like Eurostat's GESMES, like we now use WordPerfect exports from a MS-Word document. All it takes is a structured, machine readable and documented form of storage of all statistical information.

Even if all our information is available on-line, machine readable, and well documented with a lot of metadata, users will start complaining again as soon as inconsistencies between electronic publications become visible. Then, we will need some way of statistical coordination and integration, like we did with the National Accounts, but now on a larger scheme. Trying to integrate as many publications as possible in a limited number of cubicles, might be a first step into the right direction. To do so will ask for a great effort in streamlining statistical definitions and classifications. In the end, the conceptual problems might overshadow the technical ones.

Finally, there is the interesting aspect of cost and price on the Internet. Assuming that statistical information itself is a public good, statisticians are often pricing statistical information only according to the marginal costs of reproduction and distribution. Reproduction and distribution on the Internet is essentially free. So, one might wonder what to do once we are able to put *all* our gigabytes of available statistical information on the Internet. Should it be free of charge or not? This topic has resulted in lively debates at SN. So besides the technical and conceptual problems, the Internet also raises great strategically issues. Live will never be as before!

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

DATA MANAGEMENT FOR CANADA'S HEALTH INTELLIGENCE NETWORK: BUILDING A VIRTUAL INFORMATION WAREHOUSE THROUGH STANDARDS, COOPERATION AND PARTNERSHIPS

B. Bradley and J. Silins[1]

ABSTRACT

The paper discusses methods and plans for creating a virtual information warehouse to serve the needs of many interdependent organizations. Building on Health Canada's DDMS/DAIS model and experience, and its Disease Control Database, it focusses on the role of metadata and standards to enable sharing and transparent exchange and integration of seemingly fragmented resources spanning different organizations, locations, mandates, territories and jurisdictions. The importance of complementary cross-functional social networks, including common information management methods and policies, shared ownership and governance, and cooperative, consensus-based priority setting and planning, is emphasized.

KEY WORDS: Data access systems; Data warehouse; Metadata.

1. INTRODUCTION

The goal of the Health Intelligence Network is to strengthen capabilities for surveillance of population health, risk factors and disease in Canada. It is being constructed for improved monitoring of health, environmental and related socio-economic conditions, to enhance investigation and research capabilities, and to assure more effective delivery of information and knowledge, when and where needed, in support of evidence-based policy and program interventions.

2. THE ROLE OF DATA WAREHOUSING

'Data warehousing' is a popular concept in the corporate information management world. Although aspects of data warehousing have been around for many years in other guises - for example, as programs or activities associated with 'managing information as a corporate resource', developing 'data inventories and repositories', managing and delivering data in 'flattened organizational structures' for 'executive information systems', and in some organizations as 'data libraries', 'data archives', or functions for 'data dissemination, data services and data marketing' - the warehousing notion lends industrial strength to what have tended to be fragmented and sometimes fragile activities.

The concept recognises the fundamental value of data as a raw material of importance, not only for application to corporate management and strategic guidance processes, but as a commodity in the emerging information economy. It also suggests the importance of analogous mining, refining, manufacturing, shelving, inventory control, shipping, marketing, sales and delivery operations.

This recasting of emphasis appears to stem from opportunities for corporate-wide data access presented by the establishment of internal networks which link most areas of the enterprise, coupled with the ubiquitous imperatives to downsize, re-engineer, and leverage internal data and information investments. Moreover, in the ripe, technological milieu created by the new enterprise networks, the Internet has struck with lightning force.

As more and more workers and managers discover that the computers on their desks provide instantaneous connectivity not only to colleagues in all divisions and

[1] Bill Bradley and John Silins, Health Canada, Room 35 LCDC Bldg. Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0L2.

decision makers in executive suites, but to people and information in all forms and organizations all over the world, there has developed a sense that something significant is happening. And as more and more organizations recognize the power of the new technologies to reshape and improve the ways in which they do business, the interest in data and information as critical strategic and economic commodities has intensified, especially the interest in metadata.

For metadata - information *about* data and information - is the key to representing and finding what is needed on networks. Metadata are also the glue that pulls everything together in useful ways.

In the statistical arena, a central data warehousing project has been under way in the Australian Bureau of Statistics (ABS) since 1993 (Colledge and Richter, 1994). As suggested by several other presentations at this Symposium (for example, Priest, 1995; Grenier, 1995; Boucher, 1995), Statistics Canada has ongoing a number of similar systems development activities, and has been engaging actively in an agency-wide discussion of how data warehousing might improve both internal operations and revenue generation.

In the U.S., where the statistical system is quite decentralized and information products more freely available, there is strong central support from the Executive Office and Office of Management and Budget (OMB) for improving electronic access. OMB's Statistical Policy Office (1995) reports that seventeen federal agencies are already making use of the Internet to disseminate statistical information.

With the growing recognition of the importance of metadata and standards for electronic access, OMB's Geographic Data Committee is working on a standard for cultural and demographic metadata to improve the identification, access, coordination and exchange of geospatial data (Federal Geographic Data Committee, 1995). The Bureaus of the Census and Labour Statistics are working together to provide a common metadata browsing and data access tool (Capps, 1995). The Bureau of the Census is also actively attempting to define and standardise metadata elements for application across the agency (Gillman, 1994; Sundgren, *et al.*, 1995), and is working with others in an American National Standards Institute (ANSI) and International Organization for Standards (ISO) initiative to develop a standard for the description and registration of statistical data elements (ANSI X3L8, 1995).

Key elements of these statistical data warehousing activities include:

- A greater focus on data and information products from an agency-wide perspective.
- More attention to machine-readable capture of and access to metadata.
- Integration of data and metadata in a corporate information warehouse as possible, or by means of improved access systems.
- Renewed emphasis on statistical integration.
- Renewed emphasis on standardization of measurement concepts and definitions.
- Improved visibility, access, and delivery of products to external clients.
- Increased emphasis on potential for revenue generation.

The largest stakeholders are data users. University data libraries, statistical data archives, and other data user organizations have been engaging in warehousing activities for many years. Issues of metadata, which until recently was referred to as 'data documentation', have been actively and effectively tackled within these communities since the 1960's. Key actors have been the Interuniversity Consortium for Political and Social Research (ICPSR) and the U.K. Data Archive (UKDA), which are arguably the largest statistical data warehouses in the world. Other's are the International Association for Social Science Information Service and Technology (IASSIST), the International Federation of Data Organizations (IFDO), the Association of Public Data Users (APDU), the Canadian Association of Public Data Users (CAPDU), and the Council for European Social Science Data Archives (CESSDA). The extensive warehousing, preservation and metadata activities of these organizations are beyond the scope of the present discussion. An introduction to key metadata standardization activities is provided in Bradley *et al.*, (1990, 1994).

A virtual data warehousing project based on an integrated data catalogue has recently been initiated in Europe under the auspices of CESSDA, (Musgrave, 1995). Statistical metadata standards development activities have recently been undertaken by ICPSR and APDU. ICPSR has formed an international expert committee to develop a standard for microdata codebooks in the form of a document-type definition (DTD) for use with word processors and editors that support ISO's Standard Generalized Markup Language

(SGML). An APDU working group is examining the problem of codebook and questionnaire documentation for survey instruments which are prepared using computer assisted survey information collections (CASIC) methodologies (Doyle, 1994).

3. FROM DATA TO INFORMATION - KNOWLEDGE ACQUISITION IN THE VIRTUAL INFORMATION WAREHOUSE

As discussed by Hicks (1995) in the opening address to this symposium, policy workers need to be able to put data together over time from many different sectors, surveys, organizations and partners in flexible ways as demanded by the research or briefing need at hand. In the population health intelligence sector, for example, the requirement may be to show specific trends in mortality or morbidity in relation to policy and program interventions and other epidemiological, health, lifestyle, economic and social conditions; to compare different subpopulations, regions, communities or countries; or to assess the impacts of different health and social security systems, cultural milieu, environmental conditions, and approaches to promotion and prevention. The emphasis on healthy public policy - that is, on interventions in all departments and sectors that are cognizant of related health determinants and minimize health risks wherever possible - requires that the breadth of access and synthesis go well beyond traditional health organizations and data sources.

Hicks also situates statistical data and information in the broader context of knowledge development for social policy making, and stresses the breadth and depth of the data and information which are needed.

Figure 1 presents a model which the authors have found useful for discussing and refining our understanding of the processes whereby data are used to support policy development and program decision making, and for planning supporting systems. (Alexander, 1990; Bradley *et al.*, 1994). For the purposes of the present discussion, we have added to the diagram to suggest the scope of the warehouse that is needed, and to emphasize its virtual nature. Figure 2 lists some of the key functional requirements for national statistical systems that are implied by the model when coupled with observations concerning the uses of information in health and welfare policy and programming environments. (Bradley *et al.*, 1994).

As illustrated, the warehouse includes not only data, but also the information and knowledge products that are derived from them. We believe that it is not sufficient to limit the scope of the warehouse to data only. Researchers, policy advisors and decision makers seek knowledge in the first instance. When applicable knowledge does not exist, or cannot be found, attention next turns to the availability of information which can be used to create it. Drilling down to the data to create new information products is a relatively technical, time consuming and expensive process, and, in theory at least, should be a last resort[2].

The term *information warehouse* thus connotes a broader range of reusable products than just data alone. There is clearly an overlap with the traditional library whose historical role has been to house knowledge products.

Moreover, the different kinds of products in the warehouse - data, information and knowledge - should be fully integrated. How many times have statisticians, researchers or policy advisors looked at a number in a report, and wondered or been asked where it came from, how it tracks with similar measures, or why it doesn't seem to fit with understandings emanating from other research? In order to address such questions in the time frames usually available, the user needs to be able to click on the number and bring up information on the underlying, analyses, data gathering and refinement processes, as well as similar and comparable documentation, with related estimates, from all other relevant sources.

Similarly, an analyst searching microdata at the question or variable level should be able to drill up through the tables, information and knowledge products in which a variable of interest has been used, in order to ascertain and make use of the analyses that have already been performed. The new hypertext technologies make

[2] Nonetheless, provision of microdata and associated access facilities remains the highest operational priority for the warehouse, especially in its early phases. This results from the increasingly more special purpose nature of the products as one moves from data through information and knowledge, and the resulting likelihood that the available aggregates, information and knowledge products will not satisfy the requirement at hand. In the early stages of the warehouse, the information and knowledge products that have already been created from the data simply may not be easily known, nor available in a useable electronic format, so that the most expedient action for the many decision-support analysts who must respond quickly will be to create their own products without extensive reference to what has already been done. In the authors' view the capture, management and integration within the virtual warehouse of this highly distributed value added is a critically high priority.

it feasible to provide linkages from the knowledge level, through intermediate information products, to the metadata for each underlying variable, and vice versa, while SGML enables us to effect the needed standardization of products and linkages throughout the pyramid.

The virtual nature of the warehouse is illustrated by means of the dotted lines which run vertically to divide the information pyramid in thin, pie-like slices. Each slice represents a component warehouse - the warehouse of one of the many organizations that are partners on the Network.

Hence one slice might represent Health Canada's warehouse, another the warehouse at Human Resources Development Canada, a third Indian and Northern Affairs Canada, a fourth Statistics Canada , another Santé Quebec, another the Ontario Ministry of Health, another the Addiction Research Foundation, another the Centre for Disease Control in Atlanta, and so on. The list of relevant data, information and knowledge sharing partners is potentially endless.

As the authors presently envision it, the warehouse depicted in Figure 1, through its associated technology, will allow for the open-ended inclusion of component resources from all information sharing and trading partners. Partners will join in as they perceive it to be advantageous for them to do so. They will control their own warehouses, and decide what may or may not be shared. They will participate in the ownership and governance of the virtual warehouse, and agree to abide by data and information sharing policies, standards and protocols that are determined by all on a collective and mutually advantageous basis.

4. STOVEPIPES AND THE VIRTUAL INFORMATION WAREHOUSE

Priest, 1995 refers to the problem of 'stovepipe' organizations. Stovepipe organizations, especially at lower levels where the raw materials of statistical information are produced, are primarily concerned with vertical rather than horizontal flows and pressures. They are endemic to most multi-level organisational structures, and quite naturally tend to build information systems which further relatively specialised objectives and perspectives. Horizontal integration is often of lower priority.

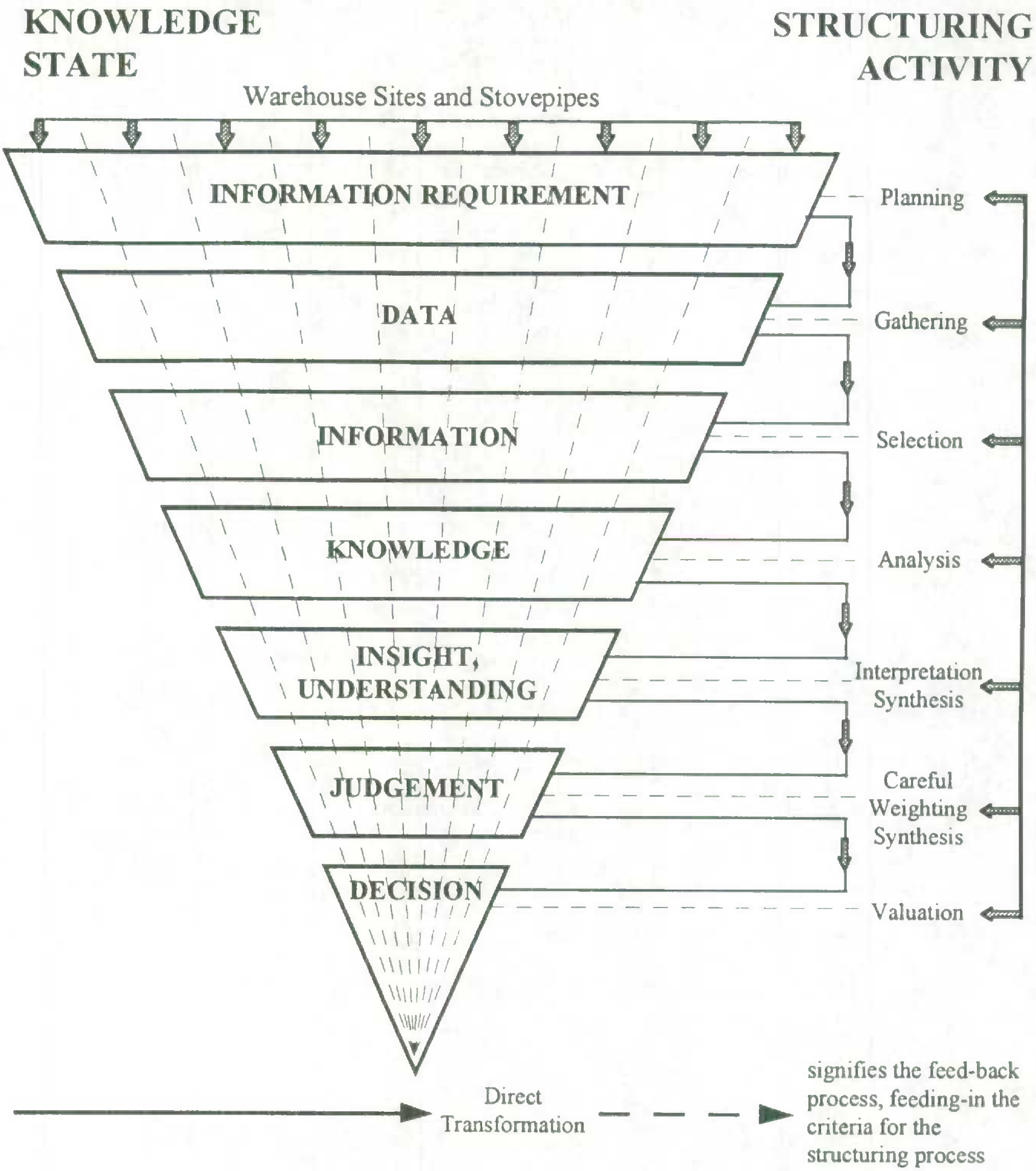
The virtual warehouse is one which facilitates the harmonization of data and information across stovepipes. Each vertical slice in Figure 1 can be viewed as a stovepipe. Depending on organizational perspective, a single slice may in fact be a collection of stovepipes. The Statistics Canada warehouse, for example, might be comprised of data, information and knowledge products from the Census, Post-Censal Surveys Program, General Social Survey, National Population Health Survey, Survey of Labour and Income Dynamics, and Special Surveys Division, to mention just a few of Statistics Canada's programs. Health Canada's warehouse might be composed of information and knowledge products from the Department's various branches, including the Health Protection Branch's Disease Control Data base, the many projects funded by the Health Programs and Promotion Branch intra and extramural research programs, and various databases compiled by the Department's Medical Services Branch. The private sector slice could encompass services from the Canada Health Monitor, Environmental Monitor, Environics Research Group, the Angus Reid Group, and other companies of interest.

The model can also incorporate public health organizations; provincial and territorial government departments; NGO's such as the Canadian Institute for Health Information, Canada Fitness and Lifestyle Research Institute, Addiction Research and Heart and Stroke Foundations; and the many university researchers, departments, institutes, archives and networks whose work contributes to population health intelligence and evidence-based policy making in Canada.

The goal is to provide a common, single window of access to the data and information products from as many of these diverse programs and services as possible. To the extent that this can be achieved, users will be able to:

- Find, browse, view and compare across all of the many servers where relevant resources are located;
- pull together all of the comparable data items and estimates points;
- carry out custom analyses, adding value by creating new information and knowledge;
- collaborate to fill gaps, avoid duplication, and improve the comparability and efficiency of future data collections;
- capture and share the value that is added at all sites and at all levels of the information pyramid.

KNOWLEDGE ACQUISITION IN THE VIRTUAL INFORMATION WAREHOUSE



Adapted from: Alexander, Cynthia: Towards The Information Edge and Beyond: Enhancing the Value of Information in Public Agencies, Justice Canada, 1990.

Figure 2.

Functional Requirements for National Information Systems

The knowledge acquisition model, Figure 1, and the authors' observations concerning the use of information in policy and program development processes in government (Bradley *et al.*, 1994), suggest the following key functional requirements for national information systems.

(I) The systems must allow access to the broadest possible base of general purpose data resources across warehouses and stovepipes at the top of the pyramid, and facilitate their retrieval, integration, refinement and delivery in the form of relatively customised information and knowledge having relevance to all organizations at the bottom.

(ii) The systems must be designed to permit movement from the top to the bottom of the pyramid, both within and between warehouses, as rapidly as possible.

(iii) The systems must allow for feedback and iteration at each stage of progression through the pyramid.

(iv) The products of all stages of the process - data, information, and knowledge - must be preserved, documented and integrated in all warehouses in a standardized, structured, accessible and reusable manner.

(v) When the available information or knowledge is questionable or found wanting, is inappropriate for the knowledge-development requirement at hand, or poses new questions, it must be possible to drill back through all levels in all warehouses from knowledge to data. It must be possible to examine and assess the suitability for any given application of all underlying resources and products, and to create new information and knowledge as necessary or desirable.

(vi) Because of the increasingly personalized, filtered, and special purpose nature of products as one moves from data, through information, to knowledge, integrated drill back capabilities founded on access to well-documented microdata are the highest priority.

(vii) It must be possible to move forward, drill-back, iterate and synthesize - i.e. to move in all directions in the pyramid, both within and between warehouses - as quickly as possible. Capability for rapid movement to and from the data level is especially important.

(viii) When gaps or deficiencies are identified in the available data sources, the drill-back must be to the planning and research design stage to ensure that relevant data are gathered or acquired as quickly as possible, that elements and definitions are harmonized across warehouses and stovepipes, and that unintended duplication is minimized.

(ix) The systems must work with local resources and warehouses in the first instance, but its scope must be global.

Improved harmonization of concepts and definitions will follow by means of a natural process of what the Australian Bureau of Statistics refers to as 'confrontation' (Colledge and Richter, 1994). By making the metadata accessible in an easily comparable way to all programs, those who plan information collections in different organizations and sectors will be able to observe, consult, and evaluate the different definitions and methods in use. They will negotiate changes in definitions from a common interest and information platform, and make adjustments as appropriate for their program and data integration objectives, research goals and client information requirements.

5. THE KEY - STANDARDIZATION[3] OF METADATA THROUGH COMMON MANAGEMENT AND ACCESS TOOLS

As discussed above, relevant data and information for policy and program analysis seldom comes from a single organization or statistical program. Even in Canada, which is thought to have a highly centralised statistical system, a single agency like Statistics Canada is only one node on the health intelligence network. There are many other important sources, including provincial governments, NGO's, universities, public health and community-based organizations. Comparable international data is also of increasing importance. The value to be captured for shared use is added everywhere, and at all levels of the information pyramid.

Program and policy analysts in Health Canada and other departments often need to tap into some or all of these sources at very short notice, to compare survey questions and other key methodological parameters across sources, and to quickly synthesize to satisfy the briefing or advisory requirement of the moment. Single-window access to ***all*** sources, through metadata that is standardized across the warehouses of the many different organizations involved, is key. The lack of such standardization is a major impediment to effective delivery of intelligence for policy and program decision making. End users in government and elsewhere have a major stake in ensuring that this kind of standardization takes place, and that it conforms to user needs and requirements.

Bradley *et al.*, (1994) demonstrate the power of standardized metadata for accessing and synthesizing cross-functional resources housed at a single site. The new information technologies make it feasible to extend the process of standardization to all program, data and information partners of relevance, so that it becomes possible to view, manipulate, analyze and synthesize the contents of all warehouses in a transparent manner, regardless of source, location, processing platform or jurisdiction.

From a technical perspective, the standardization is being accomplished by providing a common set of expert metadata creation, management, access and knowledge development tools. The tools are based on the best available standards, and widely accepted practices.

Detailed functional requirements, both for the metadata that are needed, and for associated management, standardization and access tools, are outlined in Bradley *et al.*, (1994). Many of the underlying principles and approaches have already been proven to work in the sense that they are implemented in software that has been used effectively by Health Canada and most of its data suppliers for almost a decade. (Bradley *et al.*, 1990, 1991, 1992). They are summarized below.

[3] By 'standardization of metadata' we mean standardizing the elements, machine-readable formats and structures of data dictionaries, codebooks, study descriptors, data catalogues, supporting text, and research results. At the data level these include such things as variable and value labels, field positions, question text, and skip patterns; summary descriptors of geographic universe, variable source, coverage, interpretive notes, interviewer instructions, units of analysis, target populations, and sample design; various attributes for identification and catalogue control purposes.
We are *not* referring to standardization of measurement concepts and definitions, such as the International Classification of Diseases (ICD) codes, or industy and occupation coding, which are the job of other agencies and organizations. *Our concern is with standardizing the plumbing of the systems and networks that create and deliver health intelligence - not with what flows through it.* We are dealing only with the machine-readable containers within which such codes, concepts and definitions are conveyed to users. However, as discussed above, we believe that one of the benefits of standardizing the underlying containers will be a greatly accelerated process of harmonization of concepts and definitions across statistical and information programs.

- *Documentation at Source:* The metadata should be created once and properly in a standardized manner at source. Neither user data service organizations nor their clients have the resources to prepare or redo data dictionaries and documentation in the manner needed to make all data and metadata accessible, useable and comparable - there is simply too much data, and the business of documenting it is labour intensive and costly. As discussed by Sundgren (1993), metadata elements should never be entered more than once. Documentation must be done properly at source by those who were responsible for gathering the data in the first place[4].

- *Build on Best Available Standards:* Build on the best available standards and the most widely accepted practices, especially the best practices of large data producers. Reinforce and support the systematic development of practical standards, and the convergence of standards and practices.

- *Portable Software Tools and Metadata:* The software tools for creating, editing, managing and adding value to the metadata must be portable to all data producers and users, and used by them. Similarly, the tools for finding data and information elements of relevance, for performing data extracts and analyses, and for creating, managing and accessing related information and knowledge products must also be universally accessible and useable.

- *Add Value to Stovepipe Processes:* From the viewpoint of individual survey program managers, the tools must add value to the data and information delivery activities within each stovepipe where ever possible. Standardization across programs is achieved as a by product.

- *Start with the Microdata:* Begin by creating tools for documenting microdata[5]. Aggregates are fully described by the metadata for the underlying microdata, coupled with information concerning the process(es) by which the aggregate was derived. It is not possible to understand an aggregate without first having a full appreciation of the underlying units of analysis, along with the observations made or measurements taken on them. Metadata and systems which focus on aggregates only are at risk of being incorrectly specified unless they build on those for underlying microdata.

- *Generic Data Dictionaries:* Build on generic data dictionary structures. The underlying data dictionary should be designed to work with as many statistical and information processing packages as possible, past, present and future. To the extent that this is achieved, the resulting metabase can serve as a platform for data and information access regardless of the many different software packages which may be used to carry out individual data management and information development activities.

- *Extended Codebook:* Integrate all documentation pertaining to each variable in the microdata record with the generic data dictionary elements for that variable. This results in what Doyle (1994) refers to as an 'extended codebook'.
The extended codebook contains not only computer package labels, record layout, missing value and format information which can be used on all processing platforms - the generic data dictionary elements - but also question text and source notes, sample and weighted frequencies for all response

[4] We estimate that the cost of duplication in the form of reformatting and reentering data dictionaries for different statistical packages may range up to 250 person years annually in Canada, depending upon the number of data sets released in the public domain and the numbers of users who install them. (Bradley *et al.,* 1994). This doesn't count the time that researchers spend searching for information about data sources of interest, obtaining all of the relevant documentation elements, and regenerating basic tables and information products, nor does it consider the possible duplication of metadata at various stages by those who gather and process the data in the first place.

[5] By microdata we mean data pertaining to the lowest unit of analysis, or in basic statistical terms, that which is being counted. In a survey, the microdata are the records containing observations on each individual respondent. By aggregate, we mean a summary measure or estimate, such as a count, total, mean, variance or percentage. In the context of microdata, some think of aggregates as 'macrodata' (Creecy *et al.*, 1994).

categories, subpopulation coverage notes, interpretive notes, interviewer instructions, edit and imputation rules, skip logic pointers, and other documentation elements that a user may need to consult to make use of the variable. To the extent that this is achieved, we then have a documentation schema at the variable level that can be used by all information and knowledge developers in all information warehouses and on all processing platforms in the world.

- *Generic Catalogue and Study Description:* Next layer into the structure elements of documentation which pertain to the data set as a whole. The most critical of these from a standards and structure point of view are those which library information science specialists refer to as 'bibliographic', 'catalogue' or 'study description' elements. These serve to provide identification and bibliographic control information for the data set, and to summarize its content and key scientific properties.
  The approach used at Health Canada has been to adopt a selected set of descriptors that are generic to all major bibliographic and study description systems in use, so that the elements can be mapped to all systems (Ruus, 1992). These include the Canadian, U.S. and U.K. Marc systems; Europe's Standard Study Description (SSD); and systems in use at NASA and at the Interuniversity Consortium for Political and Social Research (ICPSR).

- *Links to Supporting Text:* The dictionary, codebook and catalogue elements above are common to most microdata, form a minimal though adequate set of documentation for many common analytical applications, and can be defined and structured in a relatively straightforward and standardised manner. There is also, however, a great deal of supporting text, such as study or program background, literature reviews and rationale, papers on methodological issues, survey instruments, special studies or reports, processing summaries and flow charts, and etiologies of question or scale selection, coding manuals, and related files and correspondence. This kind of documentation must also be structured, integrated with catalogue and codebook, and made available in a portable electronic format for convenient access at the study or data set level.

- *Relational and Text Database Structures:* Health Canada has had considerable success using pseudo-relational, portable xbase structures for the dictionary, codebook and catalogue portions of the structure. The new text processing technologies, especially those based on the Standard General Markup Language (SGML), and its Hypertext Markup Language (HTML) subset, provide further opportunities for integrating, standardizing and structuring all documentation materials, for incorporating information and knowledge products, and for providing on line access across all warehouses through widely available browsers.

6. INTEGRATING INFORMATION AND KNOWLEDGE PRODUCTS IN THE METABASE

In addition to providing a complete, structured documentation package, standardized for all data sets in all warehouses, along with associated creation, management, access and extraction tools, the authors believe that it is very important to integrate information and knowledge products into the metabase. Users seek knowledge in the first instance, not data and statistics. Metadata, as we see it, are comprised not only of information and knowledge *about* data and information, but also information and knowledge *derived from* data and information.

This provision in our definition of metadata is based on purely pragmatic considerations - *on what is needed, and on what works in delivering it.* We make the case by demonstrating, as follows.

Hicks (1995) has stressed the need to show 'very long trends' across many diverse data sources to support social policy development. Consider the illustrative metadata product shown in Figure 3, which shows an important 'long trend' in the risk management area of health policy. We paraphrase Bradley *et al.*, (1994, p. 30). *This product is derived entirely from the information in ... Health Canada's metabase... With the exception of the interpretative statement given as the second bullet, each element has been pulled from the metabase, analyzed and synthesized as a coherent whole to have meaning, then reformatted for presentation.*

Metadata - elements of data documentation - are the building blocks of information and knowledge.

Figure 3 is an example of a kind of information that we call 'infobits' - fact sheets, and other stand-alone, reusable products which focus on relatively discrete and simple pieces of information. Typical infobits show the responses to a single survey question for different subpopulations or other conditions of interest, or to the same question asked in many different but comparable surveys at different points in time.

Returning to the 1994 paper, *...such products can be stored electronically in integrated fashion with the metadata, searched by automated means, and selectively retrieved, examined and printed for incorporation as supporting material in briefing notes, research summaries, and presentations.*

The resulting knowledge products, in turn, can also be stored on the metabase, where they can be retrieved for adaptation to new requirements, or for detailed scrutiny of the supporting information and data.

Following the model in Figure 1, *it then becomes possible to drill-back from knowledge to the supporting information, and from each aggregate data point in an information product to the underlying microdata documentation. ... it will be possible to click on a data point or an interpretative statement to bringup all of the supporting information and documentation. When the available knowledge or information is found wanting, or is inappropriate for the requirement at hand, the drill-back can be to the data itself to generate new aggregates and information products, or to examine the behaviour of existing aggregates when subject to different theoretical and analytic considerations.*

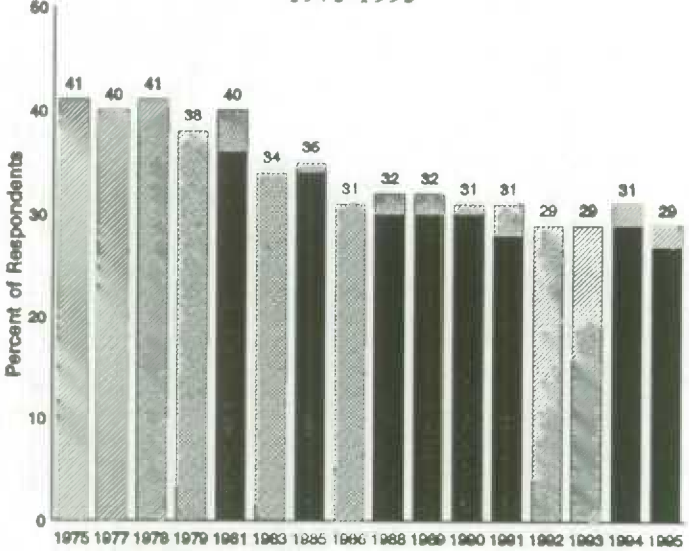
Figure 3 also illustrates one of the most important capabilities of standardized metadata ... namely their capacity to assist in integrating data and information from a wide variety of sources into a single, coherent whole. Each data point ... comes from a different survey, and the surveys were carried out by many

Figure 3. CIGARETTE SMOKERS

- *The 1975, 1977, 1979, 1981, 1983 and 1986 Labour Force Surveys of Smoking Habits, 1978 Canada Health Survey, 1981 Canada Fitness Survey, 1985 and 1991 General Social Survey, 1985 and 1990 Health Promotion Survey, 1988 Campbell Survey on Well-Being in Canada, 1989 National Alcohol and Drugs Survey, 1988-95 Canada Health Monitor and 1994-95 Survey of Smoking In Canada (Cycles 1 and 4), asked questions concerning cigarette smoking. For example:*

  At the present time, do you smoke cigarettes?
- *The results over all surveys suggest that there has been a significant decrease in the proportion of cigarette smokers in Canada since 1978. The decrease appears to have levelled off since the mid '80s.*

Do You Presently Smoke Cigarettes?
1975-1995



Percentages are based on population aged 15+ who indicated a "yes" response to the question asked. This included current regular and occasional cigarette smokers.

Same year surveys were: 40% CFSA81, 36% SMK81, 35% GSS85, 34% HPS85, 32% CHM88, 30% CSWB88, 31% CHMS90, 30% HPS90, 31% GSS91, 28% CHMW91, 29% CHMS94, 31% SSC94C1, 29% CHMS95, 27% SSC94C4.

Source: DDMS Metabase: SMK75, SMK77, CHS78, SMK79, CFSA81, SMK81, SMK83, GSS85, HPS85, SMK86, CHM88, CSW89, NAD89, CHMS90, HPS90, CHMW91, GSS91, CHMS92, CHMS93, SSC94C1, CHMS94, SSC94C4, CHMS95.

organizations in different levels of government, as well as in the NGO and private sectors. Because some of the organizations perceived themselves to be competing with others at the time, and others hardly knew of each others' work, the resulting data contain many frustrating variations in methodology and question wording which make them difficult to compare and integrate... Nonetheless, the standardized documentation makes it possible to place and manipulate the metadata and aggregates in the same logical container, and to examine the variations in a systematic manner.

Hence perhaps the most significant benefit of standardized metadata is that it becomes possible to provide tools to facilitate meta analyses across warehouse sites. As demonstrated in Bradley *et al.*, (1994) such tools will significantly improve the capabilities of knowledge workers to manufacture and manage wealth in the form of new information products, products that can be used and reused in a wide variety of information and knowledge development applications. It therefore makes a great deal of sense to integrate information and knowledge with the underlying source data by means of the metadata, and to design the metadata accordingly.

Although much work is needed to better refine our understanding of knowledge development processes, at the outset the authors see three fundamental kinds of metadata elements in the 'information' and 'knowledge' categories.

- ***Standard Tabulated Aggregates:*** Again, we paraphrase the 1994 paper. *The data points in Figure 3 are simple, univariate estimates. ... also needed are breakdowns in smoking behaviour over time by sex and age, and in view of recent changes in tax policy, especially by province. ... the addition of standard tabulated aggregates by age, sex, region, province, income, education and so forth is an easy extension to the metabase. So also are capabilities to automate the production of these and other estimates from all of the individual databases from which the individual data points must be derived, regardless of the source of the database, or the processing environment in which it is implemented.*
- ***Infobits:*** The experience at Health Canada in developing and disseminating a library of approximately 1000 infobits suggests that these low level, reusable, fact-sheet-like information products - along with software tools for creating, managing, storing, manipulating and retrieving them - will be among the most popular and valuable components in the warehouse. The experience further suggests that infobits will be manufactured from other metabase elements under factory-like conditions, and selected by clients through automated means for use in research summaries, briefing notes and other knowledge development applications. Provision can now also be made for the use of audio and visual clips as part of standardized infobit formats.
- ***High level Information and Knowledge Products:*** Following the model in Figure 1, 'High level' or 'knowledge' products are traditional hard copy publications of all kinds, briefing notes, research papers, lectures, speeches and presentations. High level products are distinguished from 'lower level' or 'information' products by the numbers of facts or 'infobits' which are synthesized, the breadth and depth of topics covered, the level of analysis, and the selective presentation and structuring of information in accordance with theoretical underpinnings or in order to focus on a specific requirement, task or application. They are usually end products in their own right, and are not often reusable as a whole for other purposes.[6]

  As predominantly text and graphics, traditional knowledge products are relatively easy to integrate in the metabase using the new hypertext structuring and standardization methodologies, especially SGML, HTML and associated editors and browsers. The emphasis at present will obviously be on traditional hard copy text publications, although emerging capabilities to use audio and video at all levels of the information pyramid suggest further

[6] While the knowledge product as a whole is not usually appropriate for other purposes, the component facts and interpretations are frequently reused in other contexts, with appropriate citations, when these happen to contribute to other knowledge development requirements. Hence the need to restructure the later as separate, reusable 'infobits'. In the modern electronic era there is perhaps nothing more wasteful than a large, hard copy compendium of statistical tables, and nothing more frustrating than to have to retype or otherwise copy a table and interpretative statements for use in some other context, if one is lucky enough to have found and accessed relevant buried treasure of this kind in the first place.

opportunities for improved effectiveness, timeliness and lower costs.

Figures 1 and 2 suggest that the priority for structure should be to provide for drill down through lower level information and estimates to the underlying metadata and data, and drill up from the variable level to the information and knowledge products in which each variable has been employed.

Drill down capabilities will allow knowledge workers to evaluate assumptions and evidence underlying resulting reports and recommendations; provide paths to information, data sources and suppliers of interest for related but different requirements; and allow quick reanalyses based on different analytic and theoretic orientations.

Drill up facilities will enable data users to see known analytic relationships and results, and to find preexisting facts and knowledge of interest to their clients or research colleagues. Drill up capabilities will also enable data planners to develop a more comprehensive understanding of the meaning and accuracy of measurements and measuring techniques than is presently possible, and to better assess gaps, priorities for measurement, and opportunities for improved harmonization across data gathering programs and research studies.

7. DATA AND INFORMATION FROM ADMINISTRATIVE SOURCES

Health Canada's experience with metadata emanates mostly from its management of microdata derived from survey data gathering processes. Important information also comes from data that are derived from administrative processes. A priority for Health Intelligence Network developmental activities is to adapt and apply similar metadata management techniques to data and information from administrative sources, so that data on outcomes can be viewed and manipulated through the same window of access as data on risk factors, determinants, and self reported health conditions.

One of our most important administrative data resources is the Disease Control Database (DCDB). The DCDB is a primary source of information on mortality, cancer, and hospital morbidity for Health Canada analysts. At its core is a huge collection of microdata that have been abstracted from hospital separation and provincial death records by provincial governments, Statistics Canada, and the Canadian Institute for Health Information (CIHI). The underlying microdata are massive, representing a census of mortality and morbidity events in Canada over many years. Like survey data access methods, the DCDB system enables the user to perform special extracts and tabulations from the microdata. However the DCDB's very large number of records and relatively small number of variables encourages a somewhat different access paradigm. Rather than carrying out frequent passes of the microdata to produce special tabulations, the modal access strategy is to integrate sets of pretabulated estimates in the database, using predetermined demographic, time and disease groupings. The estimates are then displayed and manipulated by users by means of structured query language (SQL).

Following the principles and approaches outlined above, we propose to integrate the disease control database and survey resources by means of common metadata structures. The mortality and morbidity microdata can be described and incorporated into the present browsing software for survey data in routine fashion. When standard tabulated estimates and descriptors are added to the surveys database, we will ensure that the structures can be routinely applied to tables from administrative data sources as well.

The generic data dictionary structures already in use in the present metadata management technology will permit SQL statement generation for automating access to the individual flat file components of relational structures. We propose to extend the standardized data dictionary to incorporate generic schema descriptors for complex data structures. The latter will soon be needed not only to describe administrative databases like the DCDB, but also Canada's new wave of complex longitudinal survey data. Longitudinal data from the National Population Health Survey (NPHS), Survey of Labour and Income Dynamics (SLID), and National Longitudinal Survey of Children (NLSC) will begin to appear in 1997 and 1998.

In this way the scope of the virtual warehouse will be extended from simple, flat microdata files from administrative processes, which can be implemented almost immediately, to complex files and aggregates from both surveys and administrative sources.

8. IMPLEMENTATION PLAN: INCREMENTAL DEVELOPMENT BUILDING ON WHAT WE HAVE

As discussed in Bradley *et al.*, (1990 - 1994), Health Canada and its principal survey information suppliers have been developing standardized metadata in accordance with the principals and requirements summarized above for many years. Machine readable data dictionaries were first developed in Health Canada in the early 1970's, and implemented in a standardized manner on national timesharing networks in 1975. The present activity was conceived in the early 1980's, when the PC desktop revolution had taken hold, and initiated in 1985, when it had become clear that the future lay in distributed computing integrated through local and wide area networks.

A metadata creation, management and reformatting tool known as DDMS, for *Data Documentation Management System*, was created in the mid '80's and applied by Health Canada staff to the survey data sets in use at that time. Because of the need for portability, DDMS was developed first in the ubiquitous dBASE pseudo RDBMS system, then converted to CLIPPER '86 which provided portable load modules. By the early 1990's the application was stretching the capabilities of CLIPPER '86 technology, so a supplementary system was created to add question text and catalogue descriptors to the metabase.

The present metabase describes approximately 250 survey data sets covering most major health, social, demographic, income and expenditure surveys carried out in Canada since the early '70's. The software provides an extended codebook according to the best available standards of the 1980's and early '90's, reformats its generic dictionaries as data definition programs for most major statistical packages, is being used at source by many of Health Canada's data suppliers in the public and private sectors, and has assisted data dissemination activities for universities in Canada under the auspices of data purchase consortia sponsored by the Canadian Association of Research Libraries (CARL) and the recent Data Liberation Initiative (DLI) of the Social Sciences and Humanities Research Council of Canada.

The Disease Control Database is a completed product, in ORACLE. Plans are being developed to incorporate other national datasets.

9. THE POPULATION HEALTH INTELLIGENCE DATABASE

The creation of a virtual population health intelligence database comprised of both survey and administrative data and information for use by Health Canada and it's partners is a priority goal of the Health Intelligence Network. The work will proceed in two phases.

In the first phase, to be carried out immediately, the existing DDMS and DAIS systems will be implemented for production use by all analysts on Health Canada's enterprise network. These new versions will benefit the Department's internal operations and evidence-based decision making, as well as its many cooperating data suppliers.

In the second phase, to begin early in 1996, the standardization, virtual access, integrated aggregates and information, knowledge development tools and information sharing methods embodied in the DDMS/DAIS model will be applied to and integrated with the Disease Control database in a new Windows implementation, to operate also under Unix and the World Wide Web for sharing with partners in a virtual warehouse on the Internet.

10. CONCLUSION - DATA FLOWS, UTOPIA AND THE OTTAWA ELECTRIC RAILWAY

Had the present symposium been held a half century or so earlier, most out-of-town participants would likely have arrived in Ottawa by rail and taken lodging in the venerable Chateau Laurier Hotel, which is located adjacent to Parliament Hill across from what was once Ottawa Union Station. Transportation to and from the Symposium would probably have been by means of the Ottawa Electric Railway, a rail-based system of electric powered carriages known as street cars which formed Ottawa's principal mode of public transport from the 1890's until the 1950's. Prior to electrification the carriages were drawn on rails by teams of horses.

A number of participants at the present Symposium are in fact staying at the Chateau Laurier Hotel. Imagine what it would have been like getting to and from today's meeting had each few blocks of the public transit system been controlled by separate and relatively autonomous proprietors, each with a somewhat different vision of what or whom should be conveyed, where, why, or the value of doing so, and with few if any standards concerning track size, rolling stock, and common routes, maps, schedules and fares.

Under such circumstances one might well have had to disembark a dozen times or more to change to a different system. At the entrance to each carrier's transit territory passengers would probably have had to familiarize themselves with the conveyance services offered, and the local routes and schedules. They would then probably select a route, queue up and wait for the next carriage, all the while hoping that the route chosen would link to the next carrier at a point where track and vehicles of some sort would be available to advance the traveller's progress in the desired general direction.

We would not long tolerate such a situation, nor would we consider as utopian anything other than a harmonized, metro-wide transit system, fully interoperable with regional, national and international carriers. Fortunately, even in it's horse drawn carriage days, Ottawa's rail-based public transit system did not suffer from such a lack of standards and standardization. Why then, over a century later, do we continue to tolerate such conditions in our national and international statistical information systems?

By outlining and demonstrating some of the benefits of standardizing the structure and content of metadata, we are describing a state of affairs which would be demanded as normal in domains where standardization has already been achieved for obvious practical purpose and common gain. We are proposing simply that all statistical data and value added be put on the same rails so that they can be harmonized and conveyed more effectively and meaningfully to clients, and so that all concerned can quickly realize the benefits of moving from horse drawn carriages to the electronic information highway.

The virtual information warehouse is technically and inexpensively achievable. We have provided only a glimpse of the enormous benefits that will accrue for everyone if we work together to support and achieve the needed standardization. As with widely used word processors, spreadsheets, and statistical data processing packages, the NIH (not invented here) syndrome is no longer acceptable.

While much work of a technical nature remains, the key issues are no longer technical. Primary concerns pertain to the ability of information stakeholders in all sectors to create and share common infrastructures, and to devise workable policies and procedures for managing, valuing and exchanging underlying products and resources. There can perhaps be no better initial impetus for the needed standardization efforts than improved surveillance and protection of human health and well being, but it will then become a matter for support and continued cooperation among partners.

Our experience suggests that shared ownership and governance of the resources and activities amongst all partners, and the establishment of enduring information policy and governance structures, will be critical ingredients. We invite all stakeholders to join in.

REFERENCES

Alexander, C. J., (1990). Towards the information edge and beyond: enhancing the value of information in public agencies, Report Submitted to Justice Canada, Department of Political Science, University of Western Ontario, London, Ontario.

ANSI X3L8 (1995). See, for example, Gillman, D.V. (Ed), ISO/IEC STANDARD 11179 Specification and Standardization of Data Elements, Working Draft 6, August 1995.

Balcer, M. (1992). Group facilitator presentation to final plenary, National Summit on Information Policy, Ottawa, Canada.

Beck, N. (1992). Shifting gears: thriving in the new economy, Harper Collins, Toronto.

Berk, K., and Ryan, T.(1992). Report from the Share File Committee, Statistical Computing and Graphics Newsletter, American Statistical Association.

Boucher, L. (1995). Getting to know Statistics Canada products and services the IPS way, Presentation to Symposium 95: From Data to Information - Methods and Systems, Statistics Canada.

Bradley, W. J., Diguer, J., and Ellis, R. J. E. (1990). Methods for producing interchangeable data dictionaries and documentation, Paper prepared for meetings of the International Association for Social Science Information Service and Technology (IASSIST), Poughkeepsie, N.Y.

Bradley, W. J., Diguer, J., Ellis, R. K., and Ruus, L. (1991). DDMS: A PC-Based package for managing social science data dictionaries and documentation - introduction and basic functions, 7th Draft Edition, Social Environment Information, Information Systems Directorate, Health and Welfare Canada.

Bradley, W. J. (1992). Developing information for policy decision making, Presentation to meetings of the International Association for Social Science Information Service and Technology (IASSIST), Edmonton, Alberta.

Bradley, W. J., Diguer, J., Ellis, R. K., and Ruus, L. (1992). DDMS: A PC-Based package for managing social science data dictionaries and documentation - Reference Manual, 12th Draft Edition, Social Environment Information, Policy, Planning and Information Branch, Health and Welfare Canada, April.

Bradley, W. J., Diguer, J., and Touckley, L. (1992). DDMSS: data dictionary management system supplemental - user's guide, 1st Draft Edition, Social Environment Information, Policy, Planning and Information Branch, Health and Welfare Canada, August.

Bradley, W. J., Hum, J., and Khosla, P. (1994). Metadata matters: standardizing metadata for improved management and delivery in national information systems - Parts 1 - 3, Bureau of Surveillance and Field Epidemiology, Laboratory Centre for Disease Control, Health Canada.

Capps, C. (1995). FERRET Overview - A Federal electronic research and review extraction tool for data access and dissemination of micro and macro data, Survey Modernization Programming Branch, Demographic Surveys Division, Bureau of the Census.

Cohen, D. (1993). No small change: succeeding in Canada's new economy, Macmillan Canada, Toronto.

Colledge, M., and Richter, W. (1994). Data management and the information warehouse: infrastructure for re-engineering, *Proceedings of Statistics Canada Symposium '94: Re-engineering for Statistical Agencies*.

Creecy, R.H., Gillman, D. W., and Appel, M.V. ta, (1994). Metadata, statistical software and information systems at the U.S. Census Bureau: current practice and future plans, U.S. Bureau of the Census, Washington D.C.

Deecker, G., Murray, T. S., and Ellison, J. (1993). On providing client support for machine readable data files, Household Surveys Division, Statistics Canada, Ottawa.

Dodd, S.A. (1982). Cataloguing machine-readable data files: an interpretive manual, American Library Association, Chicago.

Doyle, P. (1994). Modernizing data documentation, Agency for Health Care Policy, Presented at IASSIST ‘94, San Francisco.

Doyle, P. (1994). User documentation for CASIC systems, agency for health care policy and research, presented to APDU ‘94, Washington D.C.

Fierheller, G. (1992). Growing the infratechnology, Theme Presentation to the National Summit on Information Policy, Ottawa, Canada.

Gillman, D.W. (1994). Developing a metadata database at the Census Bureau, Statistical Research Division, Bureau of the Census, Washington D.C.

Gorman, B., Silverman, A., and McLure, J. (1992). Council for administrative renewal, presentation to Expo Innovation, Government of Canada Canadian Centre for Management Development, Ottawa, Canada.

Grenier, R. (1995). Serving clients better electronically, Presentation to Symposium '95: From Data to Information - Methods and Systems, Statistics Canada, November.

Hammer, M., and Champy, J. (1993). Reengineering the corporation: a manifesto for business revolution, Harper Collins, New York.

Hicks, P. (1992). Proposal for new data on skills and learning: social statistics for prosperity, economic union and fairness, policy, Planning and Information Branch, Health and Welfare Canada.

Hicks, P. (1995). The Role of statistics in making social policy, Presentation to Symposium '95: From Data to Information - Methods and Systems, Statistics Canada.

Ludley, J. H. (1993). The use of meta data in the UK Central Statistical Office, Information Systems, Central Statistical Office, Great George Street, London SW1P 3AQ.

MacDonald, A. (1994). Blueprint for renewal in the public service of Canada, Treasury Board Secretariat, Ottawa.

Massé, M. (1993). Partners in the management of Canada: the changing roles of government and the public service, The 1993 John L. Manion Lecture, Canadian Centre for Management Development, Ottawa.

METIS90 (1990). Users' guide to meta-information systems in statistical offices, United Nations, Geneva.

Nordbotten, S. (1993). Statistical meta-knowledge and data, Invited opening lecture, Conference Proceedings, Statistical Meta Information Systems Workshop, EUROSTAT.

Priest, G. (1995). Data integration: the view from the back of the bus, Presentation to Symposium '95: From Data to Information - Methods and Systems, Statistics Canada.

Roistacher, R. C. (1976). The data interchange file: A first report, Document No. 207, Center for Advanced Computation, University of Illinois, Urbana, Illinois 61801.

Roistacher, Richard C. (1978). A style manual for machine-readable data files and their documentation, Draft2, Center for Advanced Computation, University of Illinois, Urbana, IL 61801.

Smith, S. (1992-93). Concluding plenary address and summary, in national information summit stimulates communication and proceedings: National Summit on Information Policy, Canadian Library Association and Association pour l'avancement des sciences et des techniques de la documentation, Ottawa, Canada.

Subcommittee on Cultural and Demographic Data, (1992) Content Standard for Cultural and Demographic Data Metadata, Federal Geographic Data Committee Secretariat. USGS MS 590 National Centre, 12201 Sunrise Valley Drive, Reston VA 22092.

Subcommittee on Electronic Dissemination of Statistical Data, Electronic dissemination of statistical data, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington D.C., November 1995.

Sundgren, B. (1973). An infological approach to databases, Urval 7, Statistics Sweden, Stockholm.

Sundgren, B. (1993). Statistical metainformation systems - pragmatics, semantics, syntactics, Statistics Sweden, S-11581 Stockholm.

Sundgren, B., Gillman, D.W., and Appell, M.V. (1995). Towards a unified data and metadata system at the U.S. Bureau of the Census (BOC) (Preliminary Draft), United States Department of Commerce, Bureau of the Census, Washington D.C.

Tapscott, D., and Caston, A. (1993). Paradigm shift: the new promise of information technology, McGraw-Hill.

Zeisset, P. T. (1993). Meta-information for summary statistics: the EXTRACT experience, U.S Bureau of the Census, Economic Census Staff, Washington.

Silverman, A. (1993). Administrative renewal in the federal public service: the next generation, *Optimum,* 24-1.

Wilk, M. B. (1991). Health information for Canada 1991: Report of the National Task Force on Health Information, The National Health Information Council, Canadian Centre for Health Information, Statistics Canada.

National Task Force on Health Information project team on health policy information requirements, (1991). Health Policy Information Requirements, Canadian Centre for Health Information, Statistics Canada.

National Task Force on Health Information project team on the template, (1991). Development of a Structural Model (Template), Canadian Centre for Health Information, Statistics Canada,.

National Task Force on Health Information project team health information analysis, (1991). *Health Information Analysis: Potentials and Impediments*, Canadian Centre for Health Information, Statistics Canada.

Wolfson, M. (1992). New electronic data products - experience in Statistics Canada, paper presented at 22nd General Conference of the International Association for Research in Income and Wealth, Films, Switzerland, Analytical Studies Branch, Statistics Canada.

Office of Technology Assessment (1989). U.S. Congress, Statistical Needs for a Changing U.S. Economy, Background Paper OTA-BP-58, Washington D.C.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

GETTING TO KNOW STATISTICS CANADA'S PRODUCTS AND SERVICES: THE IPS WAY

L. Boucher[1]

ABSTRACT

IPS (Information on our Products and Services) is a Windows based search and retrieval system to assist Statistics Canada employees in identifying current information on our products and services for their clients. Primarily designed for advisory services and library staff, IPS gives "one stop" access to several metadata sources and enables organized and efficient inquiry searches. IPS is an innovative tool to facilitate the ever increasing task of finding the appropriate information for our clients from the expanding selection of products and services.

IPS is a "one-stop" access: no need to search in several places to find out what products and services are available on a given topic. It is comprehensive by including all "registered" products and services. IPS enables its users to query the thousands of products and services by words, titles, subjects and authors, among others. Once the search is completed, IPS compiles comprehensive lists of products and services. Statistics Canada staff can then request information on any, some or all items on the lists and can either download this information, provide lists to clients or assist in the creation of mini-catalogues.

KEY WORDS: Metadata; Search and retrieval system; Product registration; Electronic catalogue.

1. INTRODUCTION

Each year, Statistics Canada's advisory staff satisfy over half a million inquiries about its products and services, not to count for those inquiries directly addressed to the subject matter specialists within the organization. The task is increasingly difficult and demanding. Users and clients of Statistics Canada are looking for information ranging from very specific data needs to broad socio-economic analysis while Statistics Canada's products and services are in constant evolution.

One could imagine the typical day of one of Statistics Canada's advisory staff as the following. An inquiry is received by mail, one pulls the related printed catalogue(s) and documentation available at hand, finds the information through their index and then gets and copies the metainformation of the appropriate product or service. Often, the printed catalogue or documentation falls short of detailed information to fulfill the inquiry, the actual publication is retrieved, when available at proximity, and the table of contents is consulted and copied. The package is then ready to be sent to the client. Simple telephone inquiries would be timely processed over the phone while a more complex one would likely be processed as above.

With the volume of inquiries Statistics Canada must deal with, we needed to organize our metainformation in a centralized database that anyone could access and search, to better serve our user and client community. IPS is the solution we came up with at Statistics Canada. Instead of relying on every advisory staff and subject matter specialist to assemble their own "tools" or "shelves" along with their personal knowledge of Statistics Canada's products and services, IPS supplies access to our centralized and comprehensive metainformation.

2. IPS: A PARTNERSHIP EXPERIENCE

One would assume that IPS is the electronic incarnation of our former printed catalogue and subject specific catalogues, which became numerous, too "thick", required too much manual intervention and were already out-of-date the day they were printed. We

[1] Louis Boucher, Dissemination Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

indeed needed a smarter catalogue. In reality, IPS was originally required to fulfill an internal policy mandate.

In May 1994, Statistics Canada's senior management approved the "Policy on the registration of products and services". With the ever expanding line of products and services offerings from the various areas of the organization, Statistics Canada's management felt that we needed some coordination and consolidation in our dissemination activities. Along with guidelines on dissemination, the Policy stated that "*... all products and services will be registered in a corporate database of products and services...* ". Thereafter, such database was created in Dissemination Division and the registration process was initiated across the organization under the guidance of Marketing Division.

Is was not long before we realized the potential of such a rich database for our dissemination activities and its power to better serve our users and clients. Our Advisory Services were first contacted and helped in designing an application from the database that would help them in their daily operations, even though at this time we did not have a definite understanding of their needs, neither did they. Not long after, the Library Services joined the initiative and brought in new sets of data to be incorporated in the database along with their "catalogue perspective". Historically, the Library services has always been the "owner" of the catalogue. Several other divisions of Statistics Canada have since joined the project and contributed to its current status.

In short, IPS emerged from the opportunities created by the implementation of a corporate database rather than a singled out solution. IPS has become a front line tool for our Advisory Services, an input to the production of our catalogue and many subject specific catalogues and a general reference media for all Statistics Canada employees having to deal with inquiries.

3. METAINFORMATION AT YOUR FINGERTIPS

The purpose of this chapter is to briefly illustrate the extent of the information available through IPS and elaborate on their sources. Most of the metainformation existed prior to IPS but was rather isolated and only available for very specific needs. IPS enabled a significant breakthrough in dissemination activities by bringing together this metainformation and enabling its search and retrieval in an organized and efficient manner. You no longer have to dig for metainformation, IPS brings to its users all the updated information available from the organization. No matter where the original inquiry comes in, we have the same tool box to fully serve our clients.

In order to provide unique and timely access to this metainformation, IPS consolidated the major sources of metadata from the already existing corporate database supporting the publishing and registration processes, some library and classification files, the survey files from the "Statistical Data Documentation Systems (SDDS)" well as the CANSIM time series directory.

As stated earlier, the corporate database was first created and later expanded to enable us to design the search and retrieval application that we generally call IPS. The following arguments deal with the corporate database and its components rather than the application. The corporate database of products and services contains metainformation on all registered products and services, metainformation on articles published as well as maintaining the related identifiers for surveys and time series. These identifiers enable our users to obtain more information to better understand and use our data. Clients want information on the characteristics of the survey they come for and their availability in time series.

Metainformation on Statistics Canada's products and services represent the core of the IPS database. As stated earlier, the registration process enables continuous maintenance of a comprehensive repository of all products and services to support dissemination activities. Registration also ensures consistency in corporate "common look and feel" and pricing across Statistics Canada's products and services. Such metadata make product integration easier by helping authors and subject matter specialists in locating and consulting existing or under development products and services while fostering cooperation and minimizing duplication of efforts and offerings.

Providing information about articles published by Statistics Canada has always been a significant portion of our advisory efforts. Prior to the implementation of IPS, it was very difficult to locate and access documentation about these articles. A particular emphasis was put on metainformation for articles in the course of developing the IPS. Not only do users and clients want metainformation about our products and services, they also want to know its source.

4. THE MECHANICS BEHIND IPS

Having briefly discussed the content and sources of IPS in the previous chapter, we will now spend some time elaborating on its technical side.

IPS is a client/server system, which is designed to take advantage of the power and functionality of Windows desktop workstations, while at the same time leveraging the advantages of maintaining a central repository of Metadata about Statistics Canada's products and services, for ease of maintainability. The IPS system has three components, the files on the workstation, the files on the local LAN and the Metadata files on a central server constructed from the corporate database on products and services, the CANSIM directory and the SDDS.

The Workstation Component
The workstation contains the client part of the Fulcrum SearchServer technology, along with the ODBC drivers and some Windows system files. The amount of workstation hard drive space required is about 1.54 megs.

The local LAN Component
The local LAN contains files used to support the client part of the application itself. This includes the On-line Help files, the Census Dictionary, SIC-E files along with the files used to support the Word Wheel function of the application, as well as the executable file itself. These files (about 4.68 megs) must be maintained in a directory that all IPS operators can access (in Read Only mode, for maximum protection).

The Server Component
This central Server contains the IPS information about products and services that appears on the screen, as a result of queries entered by IPS users. It also contains the Fulcrum SearchServer Server files and all of the indexing files that allow for quick search and retrieval activities to be performed.

The only other component of the system is the communications lines. The system has been designed to be "Communication Line friendly", in order to keep the Network traffic down to the bare minimum.

5. META INFORMATION AT WORK

Most of what we have discussed so far takes place behind the scenes and IPS users see very little of it in their daily activities serving Statistics Canada's clients. Let us now have a look at their screens and explore the major functionalities now available to them via IPS. In short, IPS gives "one-stop" access to various metadata sources in Statistics Canada. Users can search the database in many ways and compile detailed lists of products and services to fulfill an inquiry. In order to facilitate these functions, several tools and helpers are made available within IPS. IPS is made of three basic screens, i.e., the main screen, the result list and the metainformation itself. Those screens are complemented by a series of pull-down menus and dialog boxes for tools and helpers.

The IPS main screen has four major components, the standard menu bar, the tool bar, the search control panel and the search score card. The standard menu bar works like any Windows based application with pull-down lists of available commands and options. An interesting option for users is the possibility to toggle between French and English without having to start a new session. This is a practical feature while answering telephone inquiries where you cannot predict the language of the client.

The tool bar offers quick, visible association to an extensive offering of commands to users and is available in all other screens. Search management tools are available for saving, retrieving and replicating earlier searches. Various output features are available throughout IPS and we will discuss them later. Reference tools like the Census dictionary and the Standard Industrial Classification are now made available for interactive consultation within IPS and plans already exist to add others. Finally, the tool bar offers extensive help through the "Guide for using IPS" button.

The search control panel enables the users to perform complex searches using Boolean logic. You can direct your search for information to words anywhere, titles, subject, authors and medium among others. The search can be performed by either keying in your search criteria or selecting your search criteria by browsing though a list (wordwheel). For example, users can search on "employment" and "unemployment" in word anywhere along with "schooling" in subject and "CD-ROM" as a medium. The search results given in the score card would satisfy all of these search criteria.

Finally, the main screen features a score card of your search. Every time you perform a search, IPS will give users an preliminary result on the number of products and services that correspond to your search. This scoring serves many purposes. First, you get an indication whether your search is fruitful or not, and this enables you to fine tune your search parameters. Secondly, it minimizes the number of transactions between your station and the database server and optimizes the data communications traffic. And lastly, you have your results by medium, i.e., you know how many printed products, articles, time series, electronic products and so on meet your search criteria.

Selecting one or all media from the score card on the main screen gets you to the second screen, the results list screen. This will display a comprehensive list of products and services from the search. The list presents the title, the identifier as well as a ranking (based of the numbers of occurrences of the search parameters) for each item. The list can be sorted by rank, title or identifier when required. The search parameters are highlighted in the titles as well. You can then select one or many items from that list.

So far, IPS has used metadata in the background of the search and has not yet shown much metainformation to the user. Selecting one or many products from the results list now gets the metainformation on the document screen. For each item selected from the result list, IPS then presents all the metainformation available about that product or service. You can then navigate within that document (next or previous term) or between documents, if you selected more than one in the list. You can copy the text to your favorite text editor if you wish. Furthermore, intuitive search is available within the body of the document text by simply highlighting a word or string of words and starting such search from there.

The document screen also gives you access to related items for that document. For example, if you had selected a printed product, the related item would direct you to electronic version, to other products and services related according to the author, survey information and time series available. The related button gives direct access to this related metadata.

Throughout these screens, IPS offers various alternatives for outputting metadata. You can send your output to a printer, create a text file or generate an SGML file. The selection for output can be by tagging documents, using current document or blocked area(s) of a document. Preformatted outputs for short and long descriptions as well as lists are available. There is also the ability to generate mini-catalogues and full catalogue as such.

6. FUTURE DEVELOPMENTS

IPS is actually used as an internal tool but will be on the Internet in the near future and will most likely made available to the public in CD-ROM version. IPS is a young corporate initiative that will benefit from future developments. Some domains of development are of particular interest to Statistics Canada.

IPS will benefit from more metainformation like a full coverage on products and services' tables of contents and articles, among others. A thesaurus is already in the works to assist users in determining the most appropriate terminology for searching. Increased access within the organization as well as outside via Internet and CD-ROM will enable clients and researchers to have a better understanding of Statistics Canada metainformation holdings as well as an incentive to extent their usage of our products, services and data.

In the near future, IPS will need to be in interaction with other metadata and "data warehousing" initiatives in Statistics Canada, like the Thematic Search Tool. Such interaction will benefit all of us and most importantly provide better insight on Statistics Canada information and services to its clients and users. IPS is a successful example of cooperation and synergy between various divisions in Statistics Canada and has and will continue to generate creative thinking and shared objectives.

SESSION 8

Electronic Information Dissemination

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

THE CRYSTAL BALL IS EXTRA: HOW TO SUCCEED IN THE MARKETPLACE WITHOUT ONE

U. de Stricker[1]

ABSTRACT

Statistical Agencies, however they perceive their role, are participants in the information business as a consequence of the value of their data. That business is characterized by significant and rapid technological change, which in turn has several significant consequences: customers demand increasing levels of convenience and flexibility in information products, and new technologies for data collection and distribution creates new competitors. One way to prosper in a challenging business is to use leverage of expertise and data in partnerships with other information industry players.

KEY WORDS: Marketplace competition; Technological change; Customer expectations.

1. THE BUSINESS OF INFORMATION

In late October, the Information Industry Association -- an influential industry group founded in Washington in 1968 -- held its annual conference in Toronto. Several speakers made delighted references to the fact that by the end of the decade, the information industry would represent a segment of the economy taking in 20% of the United States GNP.

Two things strike me: one, that 20% is a lot of GDP; two, that I am here to mention it. I am not certain that I would have been invited, with my information marketing hat on, to address the Symposium five years ago. But the world has changed.

It wasn't so long ago that public mandated statistical agencies disseminated their data primarily to those interests on whose behalf the data had been collected. Not any more. Today, statistical agencies find themselves in the information business -- whether they kicked, screamed, or were thrilled about it.

The business of information is a new reality that imposes new rules, new challenges, new opportunities, and new speakers like me at events such as this Symposium.

The business of information is a challenging one. It forces the need to harmonize revenue generation with the public good and to preserve corporate integrity while capitalizing on highly valued assets.

Then again, although you may feel the challenges of being in the information business are daunting, you have focused during the Symposium precisely on the very concerns that are central to business: meeting customer needs and expectations. If satisfying needs for a product or service -- at the right time, in the right way -- is the WHAT, then marketability is the HOW. I do not have any magic answers, but I did look in my crystal ball for some observations I hope will be helpful to any organization's efforts to establish and maintain maximum marketability. We are in a tough business, one characterized by more rapid and more profound change than anyone can remember ever happening before.

2. FIRST CRYSTAL BALL MESSAGE

The crystal ball has two messages for me. The first one is WATCH OUT FOR CUSTOMERS, AND at the same time LOOK OUT FOR (sacred) COWS. This message concerns the impact of technology on expectations among consumers of information products, and concerns your relationship with customers.

I see that the appetite for information is growing exponentially in step with the amount of information available. But that appetite is not for any old information in any old shape. Customers want

[1] Ulla de Stricker, de Stricker & Associates, 3934 Selkirk Place, Mississauga, Ontario, L5L 3L5, Tel: (905) 820-4525.

flexibility, choice, convenience, and an edge over their competitors. They want automatic delivery of information matching their interest profile; they want the ability to manipulate data exactly as they see fit, differently at different times. They are unwilling to waste time, and they demand strong application support when it is required. At times, they want interpretations and analysis; at other times, they prefer to see the basic data for themselves. They have the software tools they need to work with data, and they will get more and more as software companies respond to their requirements.

Such general traits among users of data are true, to varying degrees, across the spectrum from school children to university faculty to professionals and whatever we understand by the 'general public'. The trick, of course, will be to allow each type of customer to find the precise combination of content, delivery, manipulation ability, and price that suits him or her best. One size fits all information services will not do for tomorrow's customers, and the challenge of creating infinite custom varieties points to a need for extremely careful strategizing.

One significant consequence of increasingly stringent user demands is that their tolerance for what I call "sacred cows" is disappearing. Beyond a certain level of accuracy, reliability, and freshness, customers are not interested in the rules, regulations, and methodological ideals that underlie the data emerging from a statistical agency; they are, surprise surprise, only interested in getting their hands on information that does the job they need to get done. My industry colleagues like to say, "let the customer decide what's good enough - and good enough for what purpose".

In the context of what is good enough, we industry professionals often use an illustration they refer to as the information continuum. Customers who value instant information -- the federal budget seconds after its release, real time news feeds -- aren't bothered about formatting, typos, and the absence of editorial frills, while customers who value analysis, structure, and thoughtful commentary are usually willing to wait a few minutes or hours or months for it. Publishers typically find that they can deliver and earn money from selling information at several points along a continuum from "instant but raw" to "late but polished". The key is that customers want to choose for themselves in the full knowledge of what they are getting (for example, "this data is derived from a small sample taken 3 years ago and ..."). What is acceptable depends on the purpose being met.

What does this message mean to you, then? I believe it means this: your future depends not only on an intimate knowledge of customers' needs but also on an ongoing communication with them about the choices for delivering your data. Let me say that another way: right now, the customer typically gets to see your products when they emerge from your front door after whatever time you have spent -- sometimes months or years -- labouring over it. In the future, customers (or agents who are closer to the customers than you are) want to step inside your agency and work with you to determine the "best" points at which to extract data, and if necessary with what compromises in terms of your internal conventions.

3. SECOND CRYSTAL BALL MESSAGE

The second message is: WATCH OUT FOR SURPRISE COMPETITORS and at the same time LOOK FOR PARTNERS. This message concerns the impact of technology on the competitive position of a statistical agency or government department in the information marketplace:

It probably won't come as a surprise that any sense of security in your present market position is a false one. Technology creates a proliferation of capability that in turn creates new opportunities for organizations to participate in the information business and new choices for customers. Hence your own opportunities are also those of other organizations -- among which some are better able than you are to respond to product development time horizons that are shrinking into weeks where they used to be measured in months or years.

There are FIVE aspects to the message about technology's impact:

First, data collection and database building is no longer the exclusive domain of a select few organizations. Technology makes it possible to collect data easily and cheaply, sometimes as a byproduct of other activities in fact. For example, just as the supermarket barcode invented to ease the job of the cashiers ended up delivering valuable information about consumers' shopping habits, so too satellite technology used to track moving objects can be easily imagined in a traffic monitoring context; interactive Internet technology can perform information gathering chores conveniently and inexpensively; advances in cable and telephone technology will offer new opportunities for collecting household and business transaction data; and so on. In the absence of data readily available at a reasonable price from the obvious source, the marketplace creates another.

In this context I would like to point out a special phenomenon my US colleague Stephen Arnold calls "Network Publishing" in a forthcoming book dealing with the new information media. Enabled by the network infrastructure exemplified by the Internet, network publishing, or collaborative content creation if you wish, means that disparate nodes in a network can contribute to a shared data resource. Many information resources available over the Internet illustrate this structure: information is physically collected or created at many sites, and the network pulls it all together into a virtual entity that appears to the user as a single resource. It is easy to see how a network of contributors across the country could assemble a database between them in return for some form of compensation or other incentive (such as visibility, which in our day and age is a powerful motivator). What I am suggesting here is nothing revolutionary except in that it makes large scale efforts feasible (efforts equivalent, say, to getting every newspaper boy in the country to agree to give you a list of the makes and models of the cars parked in the driveways on their beats, a description of the bluebox contents, the number and ages of the children, the breed of the pet, etc., etc.).

A similar enabling factor is the meteoric rise of 'agent software' crawling all over the Internet looking for things. With World Wide Web content doubling every couple of dozen days, the amount of data available for mining by software agents is nothing to sneeze at.

One other element is that traditional database publishers are looking to push work off the desks of humans onto machines so as to be able to harness ever larger amounts of data and maximize productivity by applying brain power to analytical and creative tasks rather than to the drudgery of, say, indexing and classification. Text indexed by a machine may not be as perfectly indexed as it would be by an expert editor, but it is a lot faster for a human editor to work on a piece of text once the machine has had a crack at it. You have seen a similar development over the years in the area of statistical databases and won't be surprised that one of the key future areas for concern is the application of software to tasks that are now performed manually, thus freeing up time to address the challenges presented by customer demands.

The **second** aspect of the technology message is that data manipulation and distribution is no longer the exclusive domain of database organizations running dialup online services. We have the explosion of applications running over the Internet to thank for that, along with the increasing muscularity of users' machines. The TCP/IP network protocol quietly outdistanced IBM's and Novell's technologies; a piece of software created in a laboratory in Geneva in the summer of 1993 became the Mosaic/Netscape "engine" of the Internet we know today, making an applications environment out of what used to be a connectivity network. Everyone with a half decent computer and some readily available software can become a publisher and a data collector on the net, delivering information and documents whose sophisticated appearance rival anything a commercial graphics house can do, and in a fraction of the time.

The market consequence is that inasmuch as the financial realities for such "publishers" may be very different from those of large organizations, there is a downward pressure on prices and a resulting adjustment in customer expectations about prices. Perhaps I should say it less delicately: customers are critical and expect excellent value for their money, even for a very little money. Microcash may very well come to account for macro revenue.

The **third** aspect of the message is that software creates greater intimacy between sellers and buyers. For example, banks and credit card companies and their customers can be much more closely linked when the customer's computer talks to the bank's (or the airline's or the rental car company's) computer directly. The implications for building customer loyalty are obvious, as are the data collection opportunities.

The **fourth** aspect of the message is that the ability to reach customers continues to get more sophisticated. Expensive and wasted colour brochures are replaced with electronic ones updated in real time; targeting methods become ever more accurate as marketing methods reach across physical space to a virtual community of customers.

The **fifth** aspect has to do with the move from "plain connectivity" to "applications platform". Not surprisingly, the Internet is a place where that move is very strongly exemplified, and will take off seriously the moment new suites of software (such as Java and Blackbird) become widespread. For example, a Swiss concern right now offers an archive of currency exchange information, plus calculation ability, over the Internet. Tools such as Java make it possible for users of such a service to, say, request an alert service triggered by certain conditions, watch real time quotes while they are connected, make bids, etc.

The new tools will create a "commerce platform" characterized by "magnet sites" drawing together all interested participants in a single virtual domain. Let me illustrate with a neutral example from another area of interest altogether: jazz - you can make the translation into your own domain easily: say that a resource of jazz sheet music and recordings were made available in digital form over an internet site. Say that a sponsor organized a "meet and swap" forum where jazz collectors could buy and sell and generally advertise what they were looking for. Say the record labels had a storefront hooked up to the site too. Say a recording studio got into the act, offering custom recordings of otherwise inaccessible material, complete with audio sample files. Say historians added their photo and video archives of performances. Say musicologists and academic music departments across the world had their faculty and graduate students contribute their research into an ever growing archive of original content (abstracts free, full text for a fee). Say that a manufacturer of musical instruments turned up too. What about a concert program, tour schedule, and booking facility for bands specializing in jazz? What about a specialty catalogue of books and other research material? What about the museums in New Orleans? What about collaborative composition efforts? What about the tour operator organizing specialty tours to sites of historial importance for jazz buffs? What about collaborative composition of new jazz music? What about, what about .... I think you can see a mental glimpse of the opportunities for information environments dealing with your kind of data - one stop shopping taken to a fine art. Keep in mind that all this activity, both academic and commercial, takes place in virtual space thanks to the tools available to run on the Internet today.

4. THE GOOD NEWS

All this adds up to say, watch out for new competitors in places you might never have thought to look. A wide range of organizations you may never have considered to be competitors are suddenly able to gather and manipulate the data you thought of as your domain and deliver information and solutions to people you thought of as your customers. IBM, Wang, Wordstar, and even Microsoft learned that marketplace position is not durable without constant vigilance.

Is there a "good news" to go with all this? Yes there is. It is probably not lost on anyone why the next panel addresses the subject of partnerships. My industry experience -- never mind the crystal ball -- says that one way for a player in the information business to prosper is to find partners and allies to help it stay one step ahead of the technological juggernaut that could otherwise render it irrelevant; to help it leverage its information properties and its expertise; and to create mutual benefit for each other and for the customers.

Technology does not exist in a vacuum. It lives by the energy of the people who use it and develop it for their own purposes. The secret of turning technology into a success factor rather than having it be a problem is to take advantage of all the work others have done to harness technology, and to bring to the table something which can enhance the value of what it can deliver: I mean, of course, your data and the intellectual capital you have invested in it.

Funny thing: at the Information Industry Association Conference, I noticed that information company executives have that idea too.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

LANDDATA BC: PAST AND FUTURE TENSE

G. Sawayama, H.A. Kucera and E. Kenk[1]

ABSTRACT

Cybernauts extol the democratic virtues of the information highway but overlook the fact that the vast majority of us are disenfranchised by the complexity of shaking hands from our computers over the telephone. New operating systems on the horizon promise to simplify the rituals of point to point connection. Then what? How to find what you are looking for? Ordering and receiving data may become routine but making sense of what you have finally downloaded will not. LandData BC is an information service infrastructure (not a tollgate, but a traffic control system) that has been developed by Macdonald Dettwiler for the Government of British Columbia. A prototype that addresses access, distribution, and integration of spatial data (geo-graphical and textual) has been in operation since May, 1993. We have learned some valuable lessons in developing and operating the prototype that will be reflected in our technological and managerial approach as we move to a production version of LandData BC.

KEY WORDS: Spatial data; LandData BC; Information service infrastructure.

1. INFORMATION SERVICE INFRASTRUCTURES

If microwave transmitters, fibre optic cables and advanced switching define the *physical* infrastructure popularized as the information highway, then LandData BC[2] can be described as a *service* infrastructure that enables access to, and delivery of *programming* (data, applications and representations) via that highway in British Columbia (BC).

There are two aspects to launching such a service infrastructure - signing up the institutions that are sitting on large banks of that programming, and the subscribers who want it. In the mid 70's, Willis Roberts and Angus Hamilton[3] used the term information brokerage to describe this classical matching of supply and demand. In fact, there is a third aspect, and that is the cultivation of value-added services that bridge supply and demand. The challenge for information service infrastructures is to develop these components 'in-synch' with each other, and with the investment in physical infrastructure. Today Bill Gates and others stand at the same point as network television broadcasters in the early 1950's.

Information service infrastructures can fulfill two roles - one is to provide organized catalogues of data suppliers and their products; and another would be to actually deliver data from supplier to user. A well-designed information service infrastructure would provide *stubs* to allow third parties to embed the infrastructure in their value-added services. Human and machine experts would perform filtering, transformation, integration, visualization, and in some cases decision-making that delivers consumer services once-removed from the data.

These information service infrastructures will require investment that will be borne by three sources - data suppliers wanting to market their data, subscribers wanting access to data, and investors in physical infrastructure who need programming to transmit.

2. GOVERNMENT AS AN INVESTOR

It may be debatable whether market forces alone will be enough to justify the high, early investment for physical and service infrastructures. Through its commitment to the National Spatial Data Infrastructure, the current administration in the United States[4] clearly understands the symbiotic relationship between the two and the strategic importance of each.

[1] Gary Sawayama, Henry A. Kucera, and Evert Kenk, Geographic Data BC, British Columbia Ministry of Environment, Lands & Parks, 1802 Douglas Street, 4th floor, Victoria, British Columbia, Canda, V8V 1X4.

In 1989, the BC government committed itself to an information service infrastructure based on an operational need that if addressed, would have strategic market potential. The dependence of the provincial economy on natural resources, the acrimonious watershed by watershed debates over sustainable land use, and the assertion of native land claims, have been well documented.[5, 6, 7, 8]

Ad hoc approaches to accumulating decision support data were inadequate and the BC government bit the bullet to develop a land information service infrastructure that would span all land information holdings across the provincial government. There are hundreds of government departments in BC investing $50 million annually in the collection of spatial data; this figure does not include analysis and usage of that data. This land information service infrastructure, since named LandData BC, was started well before Freedom of Information and Protection of Privacy (FOI) legislation in BC, and had the modest (but formidable) objective of opening government information up to other government departments for better service delivery to the public.

LandData BC would reduce duplication of effort in collecting and managing spatial data. It would channel government decision-makers toward standard corporate information resources resulting in greater internal consistency in decisions delivered by a decentralized organization. This would address the case of waste discharge permits being issued by one branch in the Ministry of Environment, Lands & Parks, without knowledge of downstream water licenses that were being granted by another branch of that same ministry.[9]

LandData BC goes beyond the straight repository, data delivery role described earlier for information service infrastructures. It actually provides a set of expert services (without discounting the possibility of others by third parties) to deliver integrated views of data from numerous data providers. It supports the concept of data federations formed by communities of interest that strive for data standards that enable interoperability.

3. LANDDATA BC PROTOTYPE

Macdonald Dettwiler, our consultant on LandData BC, started documenting user needs in December, 1990 through a series of rapid prototypes that simulated end-user functionality. Specification followed and high level design was fixed by December, 1991. After detailed design, development and testing the prototype was delivered in May, 1993. The original aim of the prototype was to determine technical feasibility, not test business case assumptions. We have extended the prototype by retrofitting an accounting module to it so that we can operate it as an interim service infrastructure. The prototype cost just under $3 million to develop over 30 months.

LandData BC is supported by a 3 volume policy, procedure and standards framework known as the Land Information Management Framework (LIMF). Conformance to the LIMF is compulsory for BC government ministries and **in theory**, is enforced as part of the budget and expenditure approval cycles. The LIMF does not embed, but instead, references specifications such as the NAD 83 reference datum and the Spatial Archive Interchange Format (SAIF). From the outset LandData BC has been designed on the assumption that data would not be centralized in physical data stores, but would continue to be managed by the individual data custodians. The LIMF holds this virtual infrastructure together.

3.1 Status of current prototype

The current prototype is implemented on TCP/IP. The repository and accounting module are built on Sybase Open Server on a Unix host located in Victoria. User Access software has been built around a low-cost GIS, Terraview, as a Windows application and has been installed on a dozen existing PC platforms in Victoria and one regional office located 750 km away. Connectivity within Victoria is on the 100 megabit (effective partitioned bandwidth of 10 megabits) per second Metropolitan Area Network, consistent with Ethernet speeds at the clients and server. In the past month, dial-up access at 28,800 baud has also been developed. The User Access software provides for browsing the repository, electronic placement of orders and a geo-graphical user interface for determining geographical extents of spatial coverages. In the case of on-line data products discussed in Section 3.1.1, the User Access software provides a data viewer. The cost of the User Access Software bundle, consisting of Sybase and Terraview licenses and PC-NFS for communications, is $1000 but drops by 50% in quantities over 100. In selecting other products in the production version for the numbers we aim to connect, we could reach $100.

The prototype can handle up to 12 concurrent sessions (a Sybase licensing limitation that can be extended) from PC's equipped with User Access software. The repository lists 35 data sources, which in turn, catalogue over 200 land information data products that range from maps to reports to data. Electronic orders can be placed for any of the listed products from the User Access sites. Formal contracts with each of the data custodians require that they acknowledge receipt of the order within 3 days and fill it within 30 days. Depending on the physical form of the data (i.e. hardcopy, etc.), delivery is by courier or telecom. We refer to these as off-line orders.

3.1.1 On-line data products

As noted earlier, the prototype goes beyond repository and off-line delivery of data products by data custodians. Direct access through fixed queries from User Access sites is provided to three on-line data products - topography, cadastral maps and a separate registry of Crown Lands.

TABLE 1

Data Product	Archive Format	Hardware/Software Platform
Topography (graphic)	SAIF/XDR	VAX / RdB
Cadastral Maps (graphic)	Arc/Info	SUN/ESRI
Crown Registry (textual)	Database	IBM 4300/SQLDS

Fixed LandData BC queries can be issued from any of the User access sites to any of the three data products. All three data products reside on different platforms at different sites under different custodians in Victoria. The queries formulated in SQL are translated into the native query language by a Data Access Server and run against the subject data. Customized Data Access Servers were constructed for each of the three data sources to handle transactions across the open system/proprietary system interface.

The subset of the data that is spatial geometry is returned to the Data Access Server where it is translated into SAIF/XDR (at this time textual attributes are moved with SQL directly) before transmission back to the User Access site. The User Access software translates the incoming data from SAIF/XDR to Terraview to enable the user to view the data. Depending on which of the three data sources were queried and the nature of the query, the data can be viewed by the user within 10 minutes (in the case of topography) to overnight (in the case of complex cadastral map or registry queries).

More significantly, the data returned from all three sources is integrated on a common datum and rationalized against each other because what is actually happening, is a data model to data model conversion.

4. THE PRODUCTION VERSION OF LANDDATA BC (AUGUST/95 TO MARCH/97)

The production version of LandData BC will extend the number of concurrent sessions from 12 to 500 with User Access software that could be installed on PC's running Windows or NT, on Macintoshes, and on Unix platforms. The production version will enable a single query to be directed to multiple data products. For example, a single query about tenures might be parsed and directed to the Crown Land Registry, to Mineral Titles, Forest Tenure, and Agricultural Land Reserves by the query server in the production version, without burdening the end user with a stream of choices. The jury is still out on whether we will support ad hoc queries. Feedback from users suggests such an investment would only benefit a handful of *power* users while raising concerns among data custodians about carrying extra loads on their side of the firewall.

We will migrate another dozen of the more 'corporate' data products from off-line to on-line access. On-line data products must have a higher level of standards compliance and specificity to stand up to the rigours imposed by a Data Access Server.

In June of this year, we received approval of $7.4 million to proceed with Macdonald Dettwiler (MDA) in developing the production version of LandData BC. The approval falls short of the $25 million requested over 3-5 years but is not nearly as debilitating as one might expect. Sixty percent of the budget requested was to flow to potential on-line data custodians who would have to model their data in SAIF, upgrade it to meet minimum requirements (e.g. NAD 83) and perform Q/A tests on their data for internal consistency and referential integrity to guarantee integratability and interoperability in the hands of the end users. As it now stands, data custodians will have to commit their own resources to these activities, and the end result will be a protracted schedule before all of their data holdings for a given product line will be ready.

Our planned, two stage implementation of LandData BC (government first, then the public) will be modified with the advent of Freedom of Information legislation and operational considerations for the exchange of information between government and the forest industry under the new Forest Practises Code.[10] For the time being LandData BC is being developed by government and will be operated by government with the expectation that revenues will make it self-sustaining. We are following a spiral development strategy rather than a 30 month gestation for turnkey delivery.

Considering that the prototype design is over 2 years old, Macdonald Dettwiler is undertaking a technology review for us. Before we selected Sybase last time, we considered OODBMS's and hybrids, but in the end, felt them to be too risky for a quasi-operational environment. Not only do we want to revisit earlier decisions, but technology being what it is, we need to monitor developments that have taken place in the intervening period and look ahead to future architectures signalled by OGIS, OLE, SQL3 and the like.

5. A LOOK AHEAD

5.1 Evolution of internet

Our Technology Review[11] notes that Internet based services such as World Wide Web client browsers and FTP servers have matured and will reduce our development costs for the production version of LandData BC.

At the same time, Internet services have not demonstrated an ability to support spatial data viewing and querying, let alone commercial considerations such as accounting and security. There is also concern about the lack of standardization around some of these Internet services.

5.2 Emerging role of spatial database servers

At some point we all started out thinking GIS's were the technology for anything spatial, whether it be spatial data capture, analysis or management. As our data holdings increase in size and complexity many are turning from GIS's to full-featured, DBMS for the data management task. GIS's are blunt instruments for formal spatial-temporal data management. Mainstream DBMS products are growing more accommodating for spatial data.

In the case of one of our data custodians, he manages over 5200 digital topographic files in BC occupying 20 gigabytes of disk. This database will grow to 7000 files and 25 gigabytes when completely populated next year. It was being managed within a relational DBMS and even then he experienced:

☞ management problems arising from the immaturity of tools for surgical updating and backup

☞ performance problems.

Object-based search engines may alleviate the latter problem in the future.

6. A LOOK BACK

6.1 SAIF implementation in XDR

From the outset, BC's objective in developing SAIF was to define a vendor-independent exchange format for maximum portability of spatial data across proprietary technologies. Sometime in 1992, in our haste to provide access to our topographic data product as a credible test for the LandData BC prototype, we took a detour in trying to implement SAIF in XDR, a SUN product in the public domain. Our problems started when we extended XDR to make it more efficient for our purposes. By departing from the XDR standard, we cut off our long term ability to rely on the standards-based tools.

This decision has been fraught with support problems and introduced an unnecessary level of complexity for those building SAIF translators to and from proprietary GIS's.

In the past year we recanted the XDR decision and opted for a straight-forward OSN implementation using PK-ZIP. This is reflected in Release 3.2 of the SAIF Specification (URL - http://www.env.gov.bc.ca/gdbc/saif32/toc.html). In September/94 we took delivery of a SAIF toolkit (API's for C or C++ environments) that can be supplied to developers of GIS translators. By February/96 we expect to have bi-directional SAIF translators, based on this toolkit, to and from two major GIS products in wide use in BC.

SAIF XDR was implemented in the Data Access Servers for the three on-line data products in the prototype. These had to be re-written. Our topographic data was the only data currently archived in SAIF/XDR and 3600 files had to be converted to SAIF/OSN.

6.2 Repositories

Repositories fulfill both an operational (version management) and strategic role (gap analysis for program planning) for data management. The LandData BC prototype's repository serves as a catalogue of data sources and products for users of the service infrastructure. None of the products are housed within LandData BC. The repository is the *menu* by which end users place an order. The LandData BC repository also contains schema tables for accessing on-line data products.

From our experience, an information service infrastructure that only delivers an on-line index to data sources may be valued by people searching for data initially. But the value drops quickly once the end user retains and narrows down their normal avenues for searching for data. This type of use of a repository is not nearly as sustained as even a phonebook. For a service infrastructure to survive, it must offer an ability to acquire the data akin to mail-order processing.

We would go one step further in suggesting that end users who have the investment and sophistication to make on-line enquiries have an expectation of rapid response. For the production version of LandData BC, this will mean more on-line data products. In developing a prototype service infrastructure, I think we made the right decision in offering respository, ordering, on-line delivery and data integration services in our prototype. It gives potential clients a taste of full end-to-end service.[12]

In the prototype, the compilation of metadata was done by data custodians who filled out forms that were forwarded to LandData BC Administration staff for entry in the Sybase database. The same process is repeated for updates.

The production version of LandData BC must include tools for decentralized and/or semi-automated management of metadata, schema and data models by the data custodians. Even better would be an ability for the LandData BC repository to point to the native repositories. The long term role of LandData BC staff will be to audit for compliance with repository specifications and policies.

6.3 Policies

What could be more second-nature to technocrats than policies? Here is but one example.

Since 1992, the BC government has instituted a data pricing policy[13,14] that establishes the fundamental principle that data is a valuable resource, on a par with budgets and person-years in delivering government services. As such, this value is recognized by allowing ministries to place a price on their data that reflects:

- cost of compiling the data
- estimate of investment to keep that data up to date
- extent to which the data is compiled to support an organization's business mandate, or conversely, an estimate of the residual value of the data to others
- estimate of the marginal costs of data distribution.

The policy applies to all extra-ministry clients inside, and outside of the provincial government and results in the return of revenue from General Revenues to the data custodian that offsets part of the cost of data maintenance This slows down the erosion of the value of the data and provides an incentive to data custodians to maximize the residual use of their data.

This policy is important in that it is a measure of value for data in the hands of an end user, and of the effort expended by a data custodian. Moreover, the effective value of the data being transacted, is what determines the case for a service infrastructure.

Pricing policies are typical of the myriad of policy decisions that will require constant maintenance and a hint of those yet to be faced. In some cases, we will be interested bystanders to other policy forums. In Canada, our highly-regulated telecommunications policy is being **de-regulated**. This transition is a period of uncertainty for service infrastructures when it comes to questions about tariffs and data telecommunications. As in the United States, the de-regulators are shaping the competition for the physical infrastructure.

6.4 Top-down design; bottom-up implementation

Service infrastructures are long-term propositions. They require considerable amounts of engineering with potentially, years between *drawing board* and fabrication. Like any other large scale project it is vulnerable to changes in technology and moving targets arising from changes in the clients' business and technological environments.

A service infrastructure that provides access to decentralized data sources has to be able to contend with variability among its suppliers. As much as one might want to press data standards as the panacea, the reality is that unless standards are minimal, you face varying levels of compliance, or at least varying levels of ability to comply.

Our service infrastructure has to recognize that even as it builds around a given data model, its architect is already contemplating changes. Among many of our land information data custodians, their data models of today are descriptions of unmodelled legacy databases. Homogenous data models for a given dataset will be the exception, not the rule.

A spiral development will roll out services or service levels throughout the project life. This enables early realization of partial benefits. It will also imply planned review or migration of those parts delivered early against those delivered at the end of the project. The success of this strategy depends on a robust design framework.

We will have to keep the perspective that the project to deliver the production version of LandData BC will never be a shrink-wrapped, finished product. What is finally signed off is only a jumping off point for extension and re-invention. Viewed in this evolutionary context, our durability as a service infrastructure will depend on adherence to accepted architectural standards. You can follow the progress of LandData BC at http://www.lii.crl.gov.bc.ca

6.5 Data, data, data

To paraphrase the rallying cry of realtors about location, data is our lifeblood. First and foremost, this means garnering the support of data custodians. Within government, the financing of those building data inventories, comes from the same purse as those building service infrastructures. In our case, the LandData BC prototype was approved at a peak in BC's economy in 1990 from a special fund. In other words, we were not competing for budget resources with our potential data suppliers. On the road to securing approval to build the production system, the province's economic outlook had changed, the special fund had disappeared, and our data suppliers were openly critical of any plan to build infrastructure at their expense when they were being inundated with overwhelming issues around Land Use and Native Land Claims. Some conceded that they would need such an infrastructure 3 to 5 years hence, but not now. Our position was based on the lead time required to deliver three years from now, and the need to make data captured today by one arm of government re-useable by others. In the end, funding materialized for both. However, we have some ground to cover if we are to line up commitment from these same data suppliers.

The Eldorado we are all chasing is the goal of interoperability and portability - to be able to access data anywhere, from anywhere; to be able to process it anywhere; to be able integrate it with other data. People tend to think of Open Systems as standards pertaining to the hardware and software industry. Were it so, the problem may at least seem remotely solvable. Nevertheless the biggest obstacle to reaching Eldorado is data and data standards, and this is in the control of institutions that gather, manage and process enormous amounts of spatial data.

SAIF Translators only address one of the constraints to fail-safe data exchange and integration as the software 'boxes' that span proprietary GIS technologies. They require data models for the two sites exchanging data as input before we can begin passing data through them. This model-to-model translation is a key step, that is not unique to multi-vendor environments and SAIF provides for standard methodologies for modelling spatial and non-spatial data.

As noted in the table below, translators and data models will not overcome the problems of specifications compliance. If the incoming data does not conform to its specifications, then it may not perform as advertised at the user's site. Only after providing data to others do you get a true measure of its quality. As strange as this may sound, when it comes to data, error reporting by users may be considered an extension of an organization's Q/A. From our observation, specification compliance is a worthy objective with long term horizons.

Problem Aspect	Nature of Problem	Solution/Approach
Hardware/Software Interoperability	Proprietary technology	Open Systems; SAIF Translators and Data Modelling
Discipline Standards (e.g. Forest Inventory)	Multiple data definitions between exchanging agencies	Consensus of discipline experts; Federations
Specifications Compliance	Inconsistency of data from given agency with respect to its published specifications	Adherence to data specifications (including Q/A); formal data management processes to associate data to specification versions

REFERENCES

[2] Building the GIS Infrastructure in British Columbia, Sawayama, G., *Proceedings of the Canadian Conference on GIS*, March, 1992, 1014-1024.

[3] Infrastructure Information Requirements in the Maritime Provinces: An Analysis, Hamilton, A.C., Department of Surveying Engineering, U.N.B., June, 1976, 3-6.

[4] Creating a Government that Works Better and Costs Less, Report of the National Performance Review, Vice President Al Gore, 168 pp, page 116.

[5] British Columbia Employment Dependencies, final report prepared for British Columbia Forest Resources Commission, Horne, G. , and Penner, C., February, 1992, 39 pages.

[6] The Future of Our Forests, Forest Resources Commission, April, 1991, 97 pp.

[7] State of the Environment (1993). Report for British Columbia, 127 pp.

[8] Land Claims reference.

[9] Ministry of Environment Geographic Information System, Final report: Description of Physical Prototype, DMR Group Inc., May, 1990, 150 pp., pages 51-72.

[10] British Columbia Forest Practises Code, Standards with Revised Field Guide References (1994), 216 pp.

[11] LandData BC Technology Review Draft, Macdonald Dettwiler, September, 1994.

[12] LandData BC Prototype System Review Draft, Evert Kenk, Surveys and Resource Mapping Branch, BC Ministry of Environment, Lands and Parks, 59 pp., page 41.

[13] Discussion Paper: Pricing and Distribution of Digital Land Information, Forum Consulting Group for BC Ministry of Crown Lands, April, 1990, 33 pp.

[14] Ministry of Lands and Parks, and Ministry of Forests Digital Data Marketing Procedures, October, 1991, Forum Consulting Group for Ministry of Lands and Parks, 69 pp.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

SERVING CLIENTS BETTER ELECTRONICALLY: THE STATSCAN ONLINE PROJECT

R. Grenier[1]

INTRODUCTION

StatsCan Online is an online service that is controlled and operated by Statistics Canada in partnership with a private technology company.

Today StatsCan Online is accessible from any country in the world that has a Public Data Network.

In 1996 we will add access to StatsCan Online from the Internet. By that time, hopefully many of the current Internet issues will have been resolved such as: security; reliability and accessibility, etc.

OBJECTIVES

The objectives originally set for this project explain the choices we made with respect to the technology design, the data and metadata content, the marketing and the service support.

These objectives will provide lifetime goals and challenges for this service as some will never be completely satisfied.

1. To improve the accessibility, timeliness and ease of use of STC's data

    - Accessibility - 24 hours/day, 7 days/week from your office, home or hotel room.

    - Ease of use of data - not just the software interface but the more encompassing understanding of the information we generate through complete metadata and creative integration and interpretation and by referencing the subject-matter experts in STC.

2. To increase the breadth and depth of information available

    - We can afford to publish, in paper form, only a fraction of our information holdings - this reality is getting worse as costs rise - especially paper. However, with online publishing the level of detail is limited only by confidentiality considerations.

    - There will always be information published offline and its existence and how to obtain it will be described on the online service.

3. To reduce the unit cost to clients

    - Not only the actual cost of the information - wherein a subscriber pays only for the data wanted, not an entire package - but also savings to client organizations through cost-effective use of their researcher's time.

4. To improve our knowledge of client needs and interests

    - . . . and therefore the *relevance* of what we collect and the *products* and *services* we deliver.

    - This implies a very comprehensive M.I.S. as well as an electronic messaging system where clients can communicate with us and we with them. It also implies a knowledgeable sales force, a very responsive Helpline, subject-matter support and technical support.

[1] Ross Grenier, Director, StatsCan Online Project, Statistics Canada, Ottawa, Ontario, K1A 0T6.

5. To improve the cost/revenue relationship over current dissemination methods

    - Through cost reduction as well as revenue generation.

6. To design a dynamic and robust service that can be expanded and improved readily

    - To be competitive, it is necessary to be able to embrace new developments in electronic information systems.

    - Multi-media, a variety of GUI platforms, interconnectivity, wireless communications, atomic pricing, Geographic Information Systems, etc.

    - In parallel, the content as well as the functionality must also evolve to satisfy user needs.

7. To accomplish all the above without compromising the quality or timeliness of the data.

MARKETING

Marketing Strategy

- **Market Testing**

    - 9 Focus Groups
    - Telephone survey
    - Field Test with Trade Statistics (Evaluation Questionnaire)
    - Marketing Test before Launch (170 subscribers for a year)
    - Field Test - CANSIM (Evaluation Questionnaire)

- **Market Needs Analysis**

    - Target Markets - needs not products
    - M.I.S. - Track who is buying what; in what quantity; when; in what combinations, etc.
    - Client feedback electronically, through Helpline, Account Executives

Pricing Strategy

- No connect time fees in Canada
- Pay only for data you receive
- Free access to *The Daily*
- One month free if annual subscription, otherwise $25/month
- Special media access with no subscription fees
- Discount for large volume users
- Educational discount

DEMONSTRATION

A demonstration of StatsCan Online was done wherein three time series were retrieved from the CANSIM database.

SESSION 9

Panel

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

EVOLVING PARTNERSHIPS IN THE INFORMATION INDUSTRY

D. Desjardins[1]

We have a most interesting subject to discuss in this the final session of the Symposium. We also have a most distinguished group of panellists who will share their experience and perspective on the subject.

Evolving Partnerships in the Information Industry - this phenomenon of the formation of partnerships has become well established. It is not something new. It is not just a trend. It is here to stay.

A national statistical agency can easily be thought of as an information utility. Its prime mandate is to provide information on the social and economic conditions of life of the citizenry that it serves. Honouring that mandate, however does not preclude roles for other players. The usefulness of the information assets within an agency's holdings can be exploited in a variety of beneficial ways by other parties - for example, the way in which that information is delivered, the way it is packaged, and the way it is transformed into other information.

Today, our distinguished group will address this phenomenon and examine a number of its aspects. They will inform us on developments in the private sector, in the public sector, at the provincial and the municipal levels. We will also hear about developments at the international level.

There are advantages and shortcomings to these partnerships. Advantages are that information becomes more widely used, that these partnerships help foster a burgeoning information industry within Canada, and, the ultimate test, that clients are better served.

On the negative side there may be the appearance that such partnerships represent a misuse of a good obtained at public expense. There may be suspicions that licensing the private sector to use public information holdings may lead to the invasion of privacy, as in the case of direct marketing.

One issue is: what is the appropriate price for the commercial use of information collected at public expense? It should be zero, some would argue, because that information has already been paid for. Others would argue that a statistical agency has no business being in the marketplace in the first instance.

Albeit contentious, these are legitimate arguments. This morning you will hear a cross-section of views from a very informed group of panellists with a combined 100 years experience in the information industry in one capacity or another.

[1] Denis Desjardins, Statistics Canada, 10-A R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6, e-mail desjden@statcan.ca.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

PANEL PRESENTATION

J. Kestle[1]

First of all I want to say how much I appreciate being invited to be here with you today. Denis mentioned before that I have had a long career in statistics with the Ontario Government and have spent many, many days here in this room in Federal-Provincial Statistical Meetings with many of you. I certainly appreciate all that I learned through my association with Statistics Canada back in those days and being here today brings back some of those memories. As Denis mentioned I will talk about our particular company's perspective in a partnership with Statistics Canada.

First thing I want to talk to you about is what Compusearch actually does. This is a case study to give you a sense of how a private sector company uses information, some of which is provided by Statistics Canada.

The nature of our business, what we do, is provide market analysis solutions. Some people think that we sell data, but actually, mostly what we do is work side by side with marketing departments to help them answer these questions : who are my customers? where do they live? and where can I find more of them? A much smaller part of our business is actually providing raw data out there into the marketplace and most of that data is provided for use in Geographical Information System (GIS) mapping packages.

Just to give you a little bit more detail, I am going to show you a very quick sample of the kind of project that we could do for a company. This is what we call a cluster bar chart. What we do is we categorize people. It is something that would get you in a lot of trouble if you did it the way we do, but we do it for private business using statistics. What we do is we attach a label, one of sixty unique types we categorize, to every neighbourhood of Canada. And then we use that categorization to help our customers understand who their customers are. This is what we call a cluster profile for a relatively high end-product. I will reveal to you later what it is. And this is a graphical representation which is designed to show you that, of sixty segments of the Canadian population, some are over and some are under-represented in the market for this particular product.

Now in order to create this piece of information we use Census data but we also use information that we buy from the Ministry of Transportation. We use information that some of our big clients like publishers give to us about their customers. So we put a lot of data into the mix and, you folks know what cluster analysis is, that is what we do with that. What we find out is that the over-represented subgroups of the population for this particular product are Technocrats and Bureaucrats, Small City Elite, Aging Erudites, Kindergarten Boom and so on. So we have these groupings that are based on purchase behaviour, demographics and a variety of other statistics. We help our clients image their customers in those terms.

We then can turn around and link those data to other data which enable us to say that the people who buy that product are people who are, for example, likely to entertain at home two to three times a month, participate in golf, dine at private clubs, visit health clubs; and they are not really into bowling or knitting. These people like to drive Jeep Cherokees, that is, because it is a Jeep Cherokee profile. They drive Cherokees and Pathfinders and they are not really big customers for Hyundai and Geo.

This is just a very brief example to give you a sense of what we do. We do not sell raw data. We take data and we combine it with a lot of other information in order to answer those questions for our customers. And, finally, based on that profile we can put the data on a map.

Our clients are banks, insurance companies, retailers, packaged goods companies, media companies, telecommunications services, not for profit organizations, automotive manufacturers and consumer durable goods manufacturers. We also provide

[1] Jan Kestle, Compusearch, 230 Front Street, West, Suite 1100, Toronto, Ontario, M5V 3B7.

information to niche software providers, consultants and data integrators. Our firm has over 4,500 customers for these specialized data products.

The company was founded in 1974 by an entrepreneur. Sometimes we get the feeling that we are viewed as the big bad guy in this part of the information industry because we are large. But I just want to make the point that we started one person at a time. A guy by the name of Bill Goldstein, who is known to some of you, started this business out of the back of a trailer in 1974, letting retailers come to him. He wrote software so that you could put a dot on a map and draw a circle on a map. These programs could gather up the Census data and tell people who it was that was living in their trade areas. He moved that business to Toronto and was the first re-seller of the Canadian Census. The company was purchased in 1985 by the Blackburn Group of London, Ontario, a major information provider, and merged with the R.L. Polk Company, the people who do the automotive statistics around the world in 1994. Our market analysis business in North America right now is around 30 million dollars. It does not mean that we sell 30 million dollars worth of Statistics Canada data. As a matter of fact our latest estimate is that about 2% of that can be attributed to the direct re-distribution of Statistics Canada data unenhanced.

We are involved in a lot of important partnerships and, just to give you a sense, we get data from a very large number of organizations that go into the creation of those data tools that I gave you a glimpse of. But our most important partner is Statistics Canada.

The nature of our partnership with Statistics Canada as we see it is: Statistics Canada produces top quality data. You do not say it often enough. Sometimes you do, but we always say at international conferences how fortunate we are in Canada to have a data provider like Statistics Canada. We are associated with U.S. companies and companies in Europe who do not have access to the kind and the quality of data that we get from Statistics Canada. In that partnership, Statistics Canada supports these data products. They foster the professional exchange of information. Compusearch has a large staff of demographers and statisticians, large for us, about 35 or 40 people who are technical specialists. They are involved in communication on statistical and data issues with the Census Division, the Demography Division, the Geography Division, the Household Surveys Division, business data providers, the Employment Equity statistics providers, just to name a few. Our relationship with Statistics Canada, based on the content of information, and what our clients need, is wide and deep. And we appreciate the value that Statistics Canada puts on that relationship and the degree of professional exchange that they have set up with us as a private company.

And finally they are, as a supplier - and we have a lot of suppliers - they are an excellent supplier to us. They appreciate our business and we view that relationship obviously as the most important partnership that we have.

What we bring to the partnership is that we add value to Statistics Canada data. Value added to me means a number of different things. It means taking that data and putting it together with other data that is sourced outside the government or from other government departments and creating statistical and market analysis products that our customers need. But it also means sometimes slicing the data up. You know, but a lot of people do not know, how much work and how much staff it takes to give someone just the ethnic origin data for the Census Tracts in just a half hour's work. Or for a researcher to pull data, put it on a diskette, put it in the proper format and then the customer says: "Oh, I don't want it in ASCII, I want it in DBF" or "I want it in a format to bring it into Mapinfo". That is a piece of value added work that has to be done in order to ship data. The orders that we get for data in all kinds of variety of formats is something to see. And sometimes the irony of it is that it is cheaper for us to ship more data than it is to ship less. You have that problem yourselves, but a lot of people do not understand that. That is part of the value added that a company like Compusearch can bring to this partnership.

We also advocate the use of the data and we work very hard to educate users. We can play a role in identifying market needs and feed that information back to Statistics Canada. Part of our partnership is of course that we have to pay. And we pay Statistics Canada and our other data suppliers a significant amount of revenue. In summary, that is what we see as the nature of the partnership and what each of the two parties brings to the table.

The benefits of the partnership is that a company like ours can bring one-stop shopping for customers who are looking for data. One of the things I am going to talk about a little bit more this morning is how demanding customers are getting about service. Because businesses are cutting back, their marketing departments are cutting back, the old do-it-yourself approach to data development and preparation is going out the window. People are not buying data by the truckload any more and hiring people to integrate it and stitch it and sow it together. They want that to be done by somebody else.

They want that to be done by specialists who are trained because the people who do those kinds of things are expensive. So we have to bring data from a variety of sources together.

One of the great things about providing data integration is it helps provide data integrity. Because when you are putting data together from different sources it helps with your quality control process. We believe that the private sector companies in these partnerships help expand the pool of expertise. When people buy data and do not know what it is they should or should not be doing with it, then data gets a bad name, because the research turns out badly. And so part of the role of the partnership is to educate clients and that means, without disrespect to anybody's business but, that a man or woman who writes software out of their basement and takes a bunch of Census data or Compusearch data and puts it together with a piece of software may not be the best person to help a business design their market analysis program. Because, if you use the data wrongly, and you get the wrong results, then the word goes out in the market place that Statistics Canada data is garbage or Compusearch's data is garbage and it does not work.

Businesses who are going to be in the data business have to commit to training and supporting their clients in the use and the issues around the statistical data.

Private sector also has some statistical licence. We can create data that is better than marginal or 70% accurate and marketers will buy that, whereas a statistical agency tries to produce data that is as statistically correct as possible. And that is part of the role of a business in data distribution. We are responsive to the market place and able to work together with Statistics Canada to understand what the market place needs and can expand the overall demand for product.

These are issues that we face in this partnership, and things that I hope we will have some discussion about today. I think the number one issue that we face as a private company in partnership with government is the perception that the data is too expensive. And I always say when my customers tell me that data is too expensive: how much do you think it should cost? I figured out the reason why our Canadian business customers think our data, our mutual data, Statistics Canada and ours, are too expensive. It is because of their U.S. experience or what they hear about the U.S. We have a sister company in the United States who buys all of the Census and all the geographic data products from the U.S. and we pay 70 times what our sister company pays to the U.S. Bureau of the Census for one tenth the amount data. Now I happen to think that they are going to have to raise their prices because we all know how much these data cost, not only to manufacture but to package and to sell. When you realize that, then you understand why we get opposition to the cost of data.

There is a general philosophy that governments should not be raising money out of data, and, Denis mentioned, some people think "Well, you've already collected it. Why should I pay for it again?". We all know the answer to that. We pay for it again because it costs money for us to package it and re-distribute it. And governments collect data for a certain purpose and businesses want to buy it for a different purpose and there are a lot of things that need to be done for each application. So I think that cost recovery is a government policy. It is a reality. We may not like every aspect of it but it is a reality that Canada has taken a lead on. Our experience with our statistical partnerships around the world is that more governments are seeing Canada as an example. Anybody in the private sector who thinks that this is all going to be changed, that the cost recovery policies are going to go away, I think that they are missing a beat.

An issue that we have to resolve in the partnership is the perception that there is competition between the partners. I do not think that we are in competition with Statistics Canada. There is a very small part of the market place where we overlap, and where we do we need to be respectful and the best person needs to win. People have heard me say this before. We have a concern on the part of Statistics Canada employees trying to sell Census data to a business and we hear Compusearch beats Statistics Canada out of a sale. My take on that is we have been selling this data to these customers since 1974 and a lot of people will continue to buy from us because they want one-stop shopping. But the extent to which we can work together, divide the marketplace, we are very happy to co-operate with Statistics Canada to get to that.

I think a major issue in these partnerships is also what I call data leakage. We need to police agreements. Nobody is allowed to get data for free. But a lot of people think they should. A lot of people think that it is okay to take data and copy it on to several machines. These people think it is okay to take data and give it to the guy down the hall. So we have to deal with policies that control licensing, that ensure that people who buy data for the purposes of doing a project do not give it to someone else. And those are challenges that the partnerships have to face.

Future trends, to wrap up. I mentioned this earlier but our customers have become very very demanding. They want more value added, not less. It is very expensive for us as it is for you to service our clients. As I mentioned before, if research is done badly then the whole information industry gets a bad name. So it is incumbent on all of us to come up with distribution mechanisms that are not "buyer beware".

Information cannot be provided under a buyer beware scenario because it just comes back and affects your business later on. It just will not work. So, in the future, we are all going to have to do more to be sure that this data will be used properly. Because of the proliferation of PC's, everybody and their brother is in the data analysis business. Most of them do not know anything about data suppression. We get calls all the time, and this speaks to the privacy issue, "why can't I get an income number for that particular EA?" You know, and we know, there are good reasons why not. The way we provide data in this country does protect the privacy of individual Canadians. But if you take an income statistic from the Census and a household number from the Census, and you gather up the EAs in your trade area, and you create the average income, then you are going to get some pretty funny numbers if you do not understand data suppression. This point illustrates that the proliferation of data means that we have to find ways to support people more and be sure that they are using the data correctly.

We have a lot of virtual companies out there that means a lot of people have been laid off and they have been entrepreneurial and they have set up their own companies and everybody is finding their niche. So everyone wants to partner with someone else. You have companies that come together for a project, a business service kind of concept, and what that means is all kinds of complicated partnerships between data providers. So the distribution agreements, the licensing agreements, are going to have to get more complicated not less complicated.

I mentioned enterprise-wide computing already, the idea that data gets copied on all kinds of machines. The software industry have been pretty good at stopping piracy. I think that we have been pretty bad at stopping piracy of data. Frankly, I do not even know as a data provider whether we are going to be able to solve this one and wether we should. I think customers should pay for multiple copies or enterprise-wide network licensing. I can tell you that it is very hard to control.

And finally, on the future trends we just see the cost of data enhancement going up and up. People do not want to know, going back to my previous example, why they cannot add up those numbers and calculate average income. They want someone to have thought that through and provided a fix. Maybe it is not necessarily 100% accurate but it is somebody's, some methodologist's or statistician's best guess. And in order to explain, and support, and caveat, and look at the results of the research, there has to be a lot more money spent on data development. And I think that I will close there.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

PANEL PRESENTATION

D. Roy[1]

Thanks to the Symposium organizers for inviting me to participate in this discussion. Partnering is a key strategy for all of the players in the information industry.

In my remarks I'm going to summarize very quickly:

- STC experience,
- STC's view of the information market based on recent client and distributor research,
- A 'VISION' for STC dissemination and the role of partnerships,
- Types of partnerships we should pursue, and finally
- What will be our next steps in this 'brave new world'.

PARTNERSHIPS - STC'S EXPERIENCE

Today STC data are widely distributed by private and public sector partners.

The origin of this strategy was the CANSIM Time Series database which, beginning in 1976 was distributed by commercial distributors. And at that stage, may I say, we were a pioneer in online dissemination.

These were essentially non-commercial arrangements from STC's point of view. Distributors were charged only for the cost of daily updating. Beginning in 1985 with the move to 'cost recovery' a data use royalty fee was introduced.

These commercial arrangements were extended to the 1986 Census to a small number of organizations who were reselling or developing 'value added' products - notably Compusearch.

Following this learning experience and to develop a 1991 Census strategy we conducted a survey of industry practices, consulted with existing and potential distributors and developed the licensing policy and end-user agreements which are now in place. We have tried to accommodate both large and small 'value adders' and resellers. (Copies of these agreements may be obtained from the Marketing Division).

CLIENT EXPECTATIONS/INDUSTRY TRENDS

How does this line up with our clients' expectations and with trends in industry?

- The industry today is being driven by rapid technology change much of it around the INTERNET which has become the common access to online services offering dynamic or rapidly changing content. At the recent IIA meetings in Toronto presenters unanimously indicated
  - the rapid commercializing of this world
  - interwoven/complementary strategies for WWW and print/CD products
  - not a loss of clients to the 'net' but new ones.

Users expect to be able to get information as soon as they want it. Powerful search tools enable users to access information from over a million web pages. And, the ease of creating and storing information has multiplied.

What is valued is also changing. Raw data is declining in value and at the other end of the value chain clients are willing to pay a premium for fully customized services.

Clients expect to obtain data from many sources from a single supplier. They want suppliers who understand their business and want products and services tailored to their planning or reporting cycles. The Wall Street Journal now has a 'Personal Journal' which is defined by the reader. Summaries of news stories which can be expanded to link to relevant background - at the reader's choice.

Client service has become an essential element of ensuring continuing client relationships and adding new

[1] David Roy, Marketing Division, Statistics Canada, 9-A, R.H. Coats Bldg., Ottawa, Ontario, K1A 0T6.

revenues through services.

(REUTERS STORY - High cost of creating new clients)

There are large numbers of new players in the industry. Many are small firms which are highly specialized in integrating information related to a particular industry and add value in the form of trend analysis, forecasting and 'recommending'.

Partnering and strategic alliances are essential strategies which allow organizations to concentrate on what they do best. Partners share in the new value their respective strengths create together - content, technology distribution or sales' capabilities.

What do our clients want? No matter what topic or product we're researching there is consistent feedback about the following. Clients want integrated STC information from a single point of entry. They don't want to contact three, five or eight divisions. They expect common concepts and standards including geography.

Our research on CANSIM industry templates and industry profiles confirm the industry trend that clients don't want standard packaging. They want information tailored to their organization's particular needs.

Increasingly our clients also indicate that STC is only one of many sources, in particular the integrators and forecasters noted earlier. Many organizations have scaled back on staff groups like economics and research units who were acquiring and analyzing data. These services are purchased as needed - which is the stimulus that has created many of the new 'integrators and value adders'.

Recently we interviewed about 40 existing and potential distributors as part of a distribution research study. Among the findings were the views that our current practices may be limiting access to information, there are many untapped opportunities and that our business practices have not kept pace with the industry.

A VISION FOR DISSEMINATION

So given these market and industry trends how will we disseminate our information in the future and how do partnerships fit into this future?

This graphic is a visualization of a future dissemination strategy. At its centre is a corporate data warehouse of;

- unpublished summary data (IBOS)
- published data (EBOS) and it is the key to creating the ability to integrate data and information from across the Agency.

A second major element of this vision is an online interface to this warehouse, StatsCan Online, incorporating highly developed metadata and search tools. This is the key to fully exploiting our data holdings - the ability to fully inform users of the information available on a topic.

There will be four streams of output to service end users from this warehouse.

Standard Products - For the foreseeable future there will be demand for print, CD, FAX and other standard products.

Online Access - There will be both public good and commercial access to the data warehouse. 'Do it yourself' end users comfortable with our concepts and with technology will continue to exist and their numbers will grow.

As well, 'gateway' access through other online providers will expand the market to occasional users. In this stream our metadata is a key as search engines identify client specified information and then retrieve it at the client's request.

Customized Services - As the range of data in the warehouse and the metadata is enhanced the ability to provide integrated custom retrievals from published and unpublished data will increase. This will be a 'core competence' and unavailable from other sources and 'highly valued'.

Partnerships - An increasing array of relationships will evolve with information suppliers. Through these arrangements STC will be able to focus on its core competencies and to leverage other's skills to

i) integrate data with information from other sources;
ii) add value through analysis forecasting or presentation/display;
iii) provide access to markets otherwise not serviced;
iv) access product development/packaging expertise.

NEXT STEPS

Among the major initiatives towards this vision which will take place in the coming months and years are

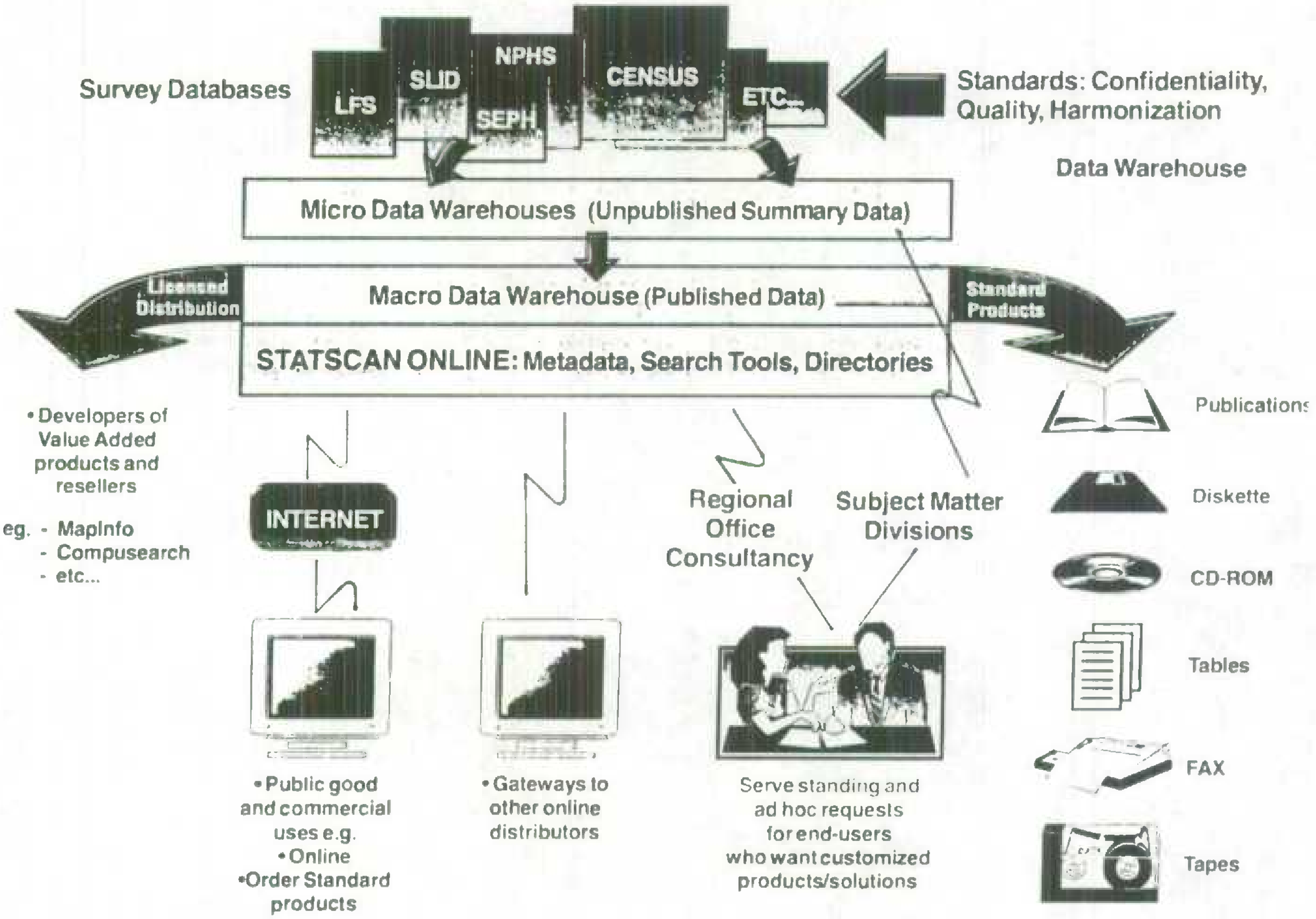
- the launch of StatsCan Online with CANSIM, Trade and the Daily;
- plans will be formulated for a multi-year CANSIM overhaul to expand it as the core of a data warehouse;

- and we expect a significant number of publications now serving small client basis will move to electronic dissemination;

- And we will continue the distributor dialogue - immediately - to understand better the opportunities for partnering, acceptable terms and the implications for the Agency.

STRATEGIC VISION
STATISTICS CANADA DISSEMINATION

Marketing Division
October 1995



Survey Databases

LFS, SLID, NPHS, SEPH, CENSUS, ETC...

Standards: Confidentiality, Quality, Harmonization

Data Warehouse

Micro Data Warehouses (Unpublished Summary Data)

Macro Data Warehouse (Published Data)

STATSCAN ONLINE: Metadata, Search Tools, Directories

Licensed Distribution

Standard Products

- Developers of Value Added products and resellers

eg. - MapInfo
- Compusearch
- etc...

INTERNET

- Public good and commercial uses e.g.
- Online
- Order Standard products

- Gateways to other online distributors

Regional Office Consultancy

Subject Matter Divisions

Serve standing and ad hoc requests for end-users who want customized products/solutions

Publications

Diskette

CD-ROM

Tables

FAX

Tapes

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

THE ALLIANCE AS AN ARRANGED MARRIAGE

A. Foster[1]

I am going to step back a little bit. Actually all this sweetness and light about strategic partnerships was worrying me a bit because the focus of my conversation is, I want to move us back to think about an analogy, the analogy of the alliance as an arranged marriage. Marriage has become quite commonplace as a way to describe how strategic alliances can work, and I think it is important for us to think about the fact that this is not a modern marriage except for the need for a clean way to get a divorce. But other than that this is not the modern romantic marriage that we have had hanging around since the 1850's. This is the arranged marriage of the 15th century for the most part and really the rules around how those marriages worked are rules that are important for people who want to make strategic alliances work. And I may need, because of the discussion today, to talk a little more about what makes them fail and about what makes them work because I did not hear enough of that in the earlier discussions. So that was not my original plan but I will try and balance this.

There are three things that are critical, that can make any of these alliances work. And if these are not in place they will not work. The first is clarity of intent: why are you getting married?

The second is the integrity of the partners. You cannot lie at the beginning of the relationship and expect it to last any amount of time. And you cannot start lying in the middle of the relationship and expect it to survive. Or, you cannot start changing the rules halfway through.

Third, your interests must be clear. And interests and intent are not the same thing and this is the importance of the arranged marriage analogy. The interests of the partners are quite separate from why you went into the relationship in the first place.

What is the essence of an intent? The intent needs to be clear. In the arranged marriage of the 15th century, the alliance took place usually between noble families or between royal houses. What they were doing in that marriage and the clear statement of intent in exchanging people and various other things but most significantly the wife, was an oath or a promise of peace. And that was the main intent, and the ultimate clarity of that intent was in exchanging people.

The intent also has to be shared and this is where it is very important that, once you have stated why you are interested in this relationship in the first place, you must be clear that you share the relationship and the intent. If, for example, the purpose of the arranged marriage was simply that somebody wanted more money and somebody else wanted peace, very often those relationships fell apart. There was a strategy on the part of a middle class families quite often to scrape up as much money as they possibly could to marry up. That created huge stresses because actually they were marrying up but the woman was marrying down and hence "disparaged" was the phrase. And this created great stresses and strain in the relationships and, by the way, in the ultimate aim of maintaining peace between countries.

Finally, the intent must be renewed on a regular basis. In this analogy people were talking about a regular exchange of gifts and it was a very calculated game. We give you a daughter, you give us some money. A little later you give my daughter some money so that she can survive if she is a widow, for example, and in between, by the way, I will give you some titles for you brother-in-law or your cousin and so on and so forth. It was a very calculated exchange.

Now what is the relevance of this understanding of intent, and how it needs to work in strategic alliances? First thing is, if you do not understand the other person's intent you are going to obfuscate their needs. If you do not understand why Statistics Canada is going into a partnership with Compusearch, or if Compusearch does not understand that, then you are not going to understand any of the decisions that take place subsequently and the

[1] Anne Foster, Carswell & Thomson Professional Publishing, Professional Bldg. 1 Corporate Plaza, 2075 Kennedy Road, Scarborough, Ontario, M1T 3V4.

grounds for misunderstanding are huge. For example what I saw being shared here were some levels of expertise. In Compusearch what you are sharing with Statistics Canada is the ability to reach the market and the ability to create value added product; and what Statistics Canada is sharing is the high quality and reliable raw data. But what is the common intent that the two have?

Next key element is integrity. And this calls for openness. There was a 14th century Lady Duborc who wrote some verses for her nephew. She was explaining to him what marriage was about and she made a very important point. She said that since her nephew's wife would be alone in the country and new to the country, she had been brought from another part of the world as part of this arrangement with no friends or family, it was his job to support her and, by the way, to state out loud what her dowry was. And that is openness and so if there is a relationship, particularly between partners who are perhaps in an unusual relationship, everyone must be open about the terms of that relationship to protect each other's interests and also to reiterate in public why you are making the arrangement.

Accurate accounting is another important part of this and the Lady Duborc spoke about that too. Having said out loud what the dowry is you have to report on a regular basis what you are doing with it until such time is as you are done. This is an important part of any strategic relationship. You have to be open about what you are doing with the money, with the resources. I did hear how Compusearch is the good guy on account of the revenues, well, let's make sure we know who is benefiting and how.

And finally, one has to offer mutual support. The interests need to be known. Now what is the difference between the interest and the intent? Well, the interests are very obviously different, it seems to me, but they need to coincide in sufficient areas so that the intent can be met.

The interest of the Government of Canada for example, and not just Statistics Canada, lies in providing for the public good through various means. The interest of the corporate partner lies in creating something and selling it for profit. Those are not mutually exclusive interests but they are different interests and so one must identify the common intent around those interests to ensure that an alliance will work.

I have a couple of examples actually if we go back to the 15th, well 16th century I guess because I am only going back as far as Henry the VIIIth, to look at strategic alliances that did not work and why. Henry the VIIIth had six wives. Of those marriages I think only two were strictly speaking arranged, in that they were intended for a larger purpose. The first one was arranged by his parents and was actually by his terms pretty successful because I think he stayed married to Catherine of Aragon for about 18 years and he could not bring himself to cut her head off. And it worked for quite a while, but Henry's interests dit not coincide with the intent. The intent was to ensure that the Pope and the kingdom of Spain were allied with the kingdom of England so that the kingdom of France did not get any bright ideas about invading or taking over little bits of property and so this was the intent. However Henry's interests became a little more earthbound and his second wife Anne Boleyn brought nothing dynastic to him, but she was not going to bring anything else either unless he married her, and so as a consequence that first dynastic marriage failed and it failed because the interests were in conflict.

The second dynastic marriage was quite a bit later. It was his marriage to Anne of Cleves. It was arranged by his Chancellor who lived just long enough to regret it, and that was intended to be an alliance of the Protestant cause at the time. Actually Henry's interest by that time had become deeply rooted in anything but alliances. And Anne of Cleves remained his wife for precisely seven days and, more power to her, she survived by understanding the interests of Henry and also understanding the intent of the factions at court. And she managed to live through this in some sort of an honourable semi-divorced state at Hampton Court, and was famous for her cooking actually. But there again the interest failed.

What are the ones that work?

Well I was going to use Catherine de' Medici as an example of a strategic alliance that worked. The reason I could not use that was that, well, she came into the partnership as an unequal partner. She was indeed just a banker's daughter from Tuscany and the royals were fairly rude to her. But her interests coincided at key points with those of various factions in the court and again she managed to sustain her interests.

So let us look at the potential partners of the marriage that we are talking about here which are the government and the private sector and let us see if we can find the shared intent. And I would argue that the shared intent amongst our two parties, at least one, there may be others but this is certainly shared, is a healthy economic environment. You cannot tax poverty. You can try but you should not. What you need is a healthy economic environment. So that is the intent in any relationship between a government and a private sector partner.

Secondly, integrity calls for two things. Do not abuse power on the government side. And you would be surprised what people view as abuse of power. This is why back there clarity is so important. In the private sector we view it as an abuse of power if we develop a software package or if we prepare a marketing plan or a marketing strategy, share it with a potential government partner, lose the bid - that is okay that is not abuse of power - but if we lose the bid and a year later up out of some government department, or indeed one of our competitors, suddenly comes this same surprising package, and this has occurred, then that is an abuse of power.

Using the Access to Information Act or using copyright as a way to exclude access to information: that is an abuse of power and therefore the integrity of the relationship is at risk. On the other side, the private sector has to know and respect the skills and the value of what occurs in government. I mean at one stage in my life I was with the Canadian Law Information Council up here. I am sure we have all been in the room when you see the curling lip of the private sector person sort of dismissing some of the efforts and the investment that has been made in creating products, or people just simply underestimating the effort required to do what we are asking from the private sector side. I was going to argue that if we in the private sector had more women in senior positions there would be less of this, but Jan and I are ruining the sample here.

The interests I touched on earlier are also different. The interests of government are around dissemination for the public good and the interests of the private sector are around publication for profit. That is not a problem; it is simply that we need to understand it.

Leon Battista Alberti was a merchant in Florence in the 15th century. Many Florentine families kept records of their family history; and marriage took up 50 - 60% of the sort of business planning time in those families so they reflected on it a lot. He said "many marriages have been causes of a family's ruin because concluded with quarrelsome, litigious, proud or malevolent individuals". And I would argue that this is true of strategic partners as well and on the business side what happens if it does not work is that you lose control. You lose control of your product or you lose control of the information that comes out of the government, you lose political control and life becomes complicated.

Results of a marriage that works, however, are new possibilities, new contacts, new sources of information and new allies. And in all the change that we have been talking about all of us need new allies. Because we cannot succeed to build what I argued was our common intent, the strong economic environment, without it. Thank you.

Proceedings of Statistics Canada Symposium 95
From Data to Information - Methods and Systems
November 1995

PARTNERSHIPS IN THE INFORMATION INDUSTRY, OR TEAMING FOR CROQUET IN WONDERLAND

P. Brandon[1]

Good morning!

I've been asked to speak about partnerships in the information industry. Being the editor and publisher of a rag called "Electronic Information Partnerships," somehow the assumption is made that I must be some authority in the area of partnerships. Not wanting to disappoint, I didn't dispute the assumption. So here I am...

And before you find out how little I know about the subject, I want to make sure we have a shared sense -- a shared context and a few shared metaphors -- of the information industry and information business in this day and age. So let me share with you the mental picture I have of the information industry and the information business in general. For that, I was irresistibly drawn to Lewis Caroll.

Remember the croquet game in *Alice in Wonderland*? In that fictional game nothing remains stable for very long. Everything is alive and changing around the player. The mallet Alice uses is a flamingo, which lifts its head just as Alice tries to hit the ball. The ball, in turn, is a hedgehog, another creature with a mind of its own. Instead of lying there waiting for Alice to hit, the hedgehog unrolls, gets up, moves to another part of the court, and sits down again. The wickets, if you remember, are card soldiers, controlled by the Queen of Hearts, who constantly changes the structure of the game by asking the wickets to reposition themselves around the court.

Now, we come to the actualization part

You are free to actualize the game using real players of your own choice. By way of an example, here's how a Statistics Canada employee might be tempted to do this. For that, I'd like to substitute *the average Statistics Canada person in the information business* for *Alice* the croquet player, *information technology* (IT) for *mallet*, *information* for the *hedgehog*, *the expectations on the employee* for the *wickets*, and the *Chief Statistician* for the *Queen of Hearts*, and I submit to you that the analogy may very nicely fit the contours of virtually everybody's predicament today!

Again, you are encouraged to come up with the substitution of your own. In fact I would like to hear about some of the more wicked actualizations, and I promise to publish them and give the authors due credit.

On a serious note, I think that *Croquet in Wonderland* -- or perhaps *Croquet in Cyberland* -- is not a bad metaphor for the kind of environment, challenges, context and wacky rules those of us in the information business face these days. I am sure I can convince my co-panelists that each and everyone of them could describe their daily existence and travails in terms that approximate the pains, joys and challenges of any Alice playing the *jeu du jour*: Croquet in Cyberland.

So, having thus set the metaphorical context for my inquiry, let me recontextualize the question. What do partnerships mean in our information age croquet game? What are some of partnership rules, if any, in this croquet game of ours.

So let me try and give you, in almost David Letterman style, ten short propositions I see emerging as rules for croquet partnerships.

I'll start with number 10, and move up to number 1.

PROPOSITION #10: We are going to have to learn that one size does not fit all. We will need to learn to structure specialized partnerships: governance partnerships, executing partnerships, advisory partnerships.

Each of them is different; it has a different "genetic code," different priorities, different focus, different rules.

[1] Peter Brandon, Partner, Sysnovators Ltd., and Editor & Publisher, Electronic Information Partnerships, 17 Taunton Place, Gloucester, Ontario, K1J 7J7, email: pbrandon@fox.nstn.ns.ca.

PROPOSITION #9: We need to re-invent distribution partnerships in a world where distribution costs are fast approaching zero.

This is a reality brutally intruding in the information world today. As information becomes increasingly liquefied, turning atoms (of paper or silicon) into bits, distribution costs and opportunities to make a living as part of a distribution chains are shrinking. I believe we need to redraw the value chain between the information supplier and the user, and totally re-map the space heretofore labelled "distribution" or "distribution chain".

PROPOSITION #8: We will need to learn to do partnerships in a world of increasing returns.

We know how to do them in a world of zero-sum moves and diminishing returns. We need to learn the art of doing them under increasing returns -- indeed, what our information economy seems to determined to confront us with. In this world, our long-held economic tenets and related beliefs (scarcity drives demand and decreasing returns, in particular) are being turned upside-down. In the words of one enlightened mind among us:

> "With physical goods, there is a direct correlation between scarcity and value. Gold is more valuable than wheat, even though you can't eat it. While this is not always the case, the situation with information is often precisely the reverse. Most soft goods increase in value as they become more common. Familiarity is an important asset in the world of information. It may often be true that the best way to raise demand for your product is to give it away." -- **John Perry Barlow, lyricist, retired cattle rancher and co-founder of the Electronic Frontier Foundation, in *The Economy of Ideas*, Wired Magazine, March 1994.**

In the world of objects, familiarity breeds contempt in the way of decreasing returns. In the world of information, familiarity breeds demand in the way of increasing returns. Software companies see demand going up after they give a batch away. Go figure!

PROPOSITION #7: Partnerships will have to go right down to and tap the core strengths of the partners.

Here's an interesting example of just such a partnerships, in which each of the partners brings to the alliance what they are really good at. The Hong Kong Trade Development Council has, in the words of its chairman, "become the de facto spearhead of a lot of Hong Kong overseas promotional efforts." The council's mandate is to promote trade around the world in order to diversify export markets; to upgrade product design and image through branding; to make Hong Kong the trade and exhibition centre of Asia; and to maintain an environment around the world for Hong Kong products to flourish. The council's efforts are well financed -- its 1992 budget was HK$784 million (US$100 million). *58% of the total income (HK$419 million) was received from its share of a government-imposed levy of .05% of Hong Kong imports and exports*, while the rest was generated through its own efforts.

So, here's an example in which the government's core competency in the area of revenue collection through taxation is partnered with an organization's unique ability to deliver a strategic information service. The result is that each party does what they are good at, with beneficial effects on the partnerships.

PROPOSITION #6: Partnerships will have to be based increasingly on cultural compatibility, trust and ethical codes and principles, rather than exclusively on contracts and legal agreements.

Traditional legal notions and concepts may become difficult to uphold in the electronic age. New forms of intellectual property may emerge. Ethics may selectively replace the law and formal regulation, perhaps not unlike the code of the Old West, which stood in for an otherwise unenforceable law. Writes John Perry Barlow in his 1994 *The Economy of Ideas* article in *Wired* magazine,

> "Until the West was fully settled and "civilized" in this century, order was established according to an unwritten Code of the West, which had the fluidity of common law rather than the rigidity of statutes. Ethics were more important than rules. Understandings were preferred over laws, which were, in any event, largely unenforceable."

PROPOSITION #5: Information technology is favouring external coordination through markets, rather than internal coordination within firms. This will lead to companies "buying" more than "making". In turn, this will lead to more use of partnerships.

As technology makes it easier, faster and cheaper to coordinate information, firms are discovering that it can be more efficient and flexible to "outsource" for goods and services rather than generate them from within a firm. Two MIT economists, Thomas Malone and John Rockart, argue that vertical integration within an enterprise is increasingly less efficient as it becomes feasible to rely on smaller, market-sensitive firms.

As a result of this trend, there may be a greater blurring of the lines that currently separate internal and external business coordination. Inter-business relations may become more collaborative than competitive as information technologies make it both feasible and attractive for firms to pursue long-term strategic alliances.

Peter Drucker compares the emerging business structure to an orchestra or hospital: top managers directly supervise a range of largely autonomous specialists who "direct and discipline their own performance through organized feedback from colleagues, customers and headquarters." He calls this business structure "an information-based organization." Through partnerships, we will extend the walls of our information-based organizations. We will need to learn how.

PROPOSITION #4: Partnerships will require increasingly more and faster sharing of information, of values and expectations.

One notion which seems to be increasingly popular these days is the idea of the "information continuum". Increasingly distributed organizations, with growing numbers of contractors, suppliers and partners, are discovering that the only thing that ultimately glues these players together is information. It is this information space, continuous, shared and eminently available, which guarantees coordination in efforts, consistency in aims and synchronization in delivery.

In hockey terms, the information continuum -- this shared information space between partners -- can be viewed as the skating ice: smooth, no cracks, no discontinuities, allowing the players to glide without fear, the puck to move without interference from the ice, the players to focus on the game. In croquet terms, you can view the information continuum as an excellently maintained, closely cut patch of quality grass, which facilitates the playing of the game.

So, for partnerships to work in our game of cyber-croquet, we must attend continuously to the quality of the playing field -- the information continuum -- which guarantees predictability and level playing conditions for all players involved.

PROPOSITION # 3: The playing field -- the information continuum -- has special needs. Its maintenance entails high transaction costs, undersupply of fertilizer (metadata) and lack of shared contexts that plague many otherwise worthy associations.

Some of the realities which will require us to be aware of and particularly vigilant as we build and maintain our informational playing fields include:

- the fact that, intrinsically, information has very high transactions costs associated with it. The cost of finding the right information, at the right time and having it delivered at the right place and in the right format is still unreasonably high;
- that there is chronic undersupply of metadata -- information about information. This, of course, increases the search costs. The problem is that there is little glory and probability of making money from metadata. One direct corollary of the chronic undersupply of metadata: there are few labelling standards in the world of information, and indeed few reliable labels on many of the information objects crossing our electronic spaces;
- that information only makes sense in a particular context. Take it out of context, and it will lose its meaning. Place it into a different context, and it will mean something entirely different. We often, groundlessly, assume that a shared context exist. Besides, we are not particularly adept at manufacturing context, and ensuring that that context actually travels with our messages. As a result, our messages are often taken out of context -- or in the wrong context -- and result in undesirable outcomes.

So what do we do? Well, for one thing, we will need players who are in the business of helping us establish those easily accessible shared information spaces, those smooth and predictable playing fields. Electronic brokers -- such as electronic commerce intermediaries or value-added network services providers -- are one category of "greenskeepers" who will probably be able to help us here. However, these players will need legitimacy and a legal and regulatory framework for them to operate, depending on the security and trustworthiness of the information continuum required to be established and maintained. The information greenskeepers' job will be a challenging one indeed.

PROPOSITION #2: We are going to have to learn the meaning of *subsidiarity* in structuring partnerships.

Subsidiarity is the key principle behind the Jacques Delors concept of the European Community (EC). In an

EC context, it simply means that the power resides within the individual countries in the Community. Only with their agreement can Brussels exercise any authority. So in a way, subsidiarity is the reverse of empowerment. It is not the centre giving away or delegating power; instead, power is assumed to lie at the lowest point in the hierarchy. The Catholic church, for instance, works on this premise when it says that every priest is a pope in his own parish. (So here I am now, in my desperation to persuade, bringing even the Catholic church into this!)

What does subsidiarity mean for partnerships? It means that the partners aren't merely cogs in the wheels; they actually have power and say and real responsibility and ability to do things not just by virtue of being your partner: they have these things inherently, by virtue of being who they are.

On the croquet field, subsidiarity means that the team draws its legitimacy from the individual talents, concerted efforts and willingness of the players to play together. The team is the functioning partnership. (Forgive me if I sometimes forget that croquet is not a really a team sport. I must confess, forgetfulness is a convenience which allows me to occasionally stretch my metaphors beyond their natural elasticity.)

And now, my final proposition. PROPOSITION #1: We are going to see the intrusion of political principles on the croquet field: the croquet field is becoming a federated space, a space of relationships in which we finally get a grip on what the truly scarce resources are.

We are brainwashed into believing that the scarce resources are things like spectrum, bandwidth of our conduits or the speed of computers. Technology's march over the last 30 years is really the story of that idea becoming unglued. We have learned to turn those things into virtually limitless commodities.

The really scarce resources, however, remain those affecting human relationships:

(1) attention -- we have not yet found ways to give people a surfeit of attention, so they can listen more intently for longer periods of times, and absorb more;

(2) "people bandwidth" -- did you know that we communicate (i.e., exchange information) with each other at the pitiful rate of 55 bits per second? (by way of comparison, PCs exchange information at 10 million bits per second, and soon they will be conversing at 155 million bits per second!); Having some personal familiarity with a person to whom you are sending a fax or e-mail message can greatly enhance the quality of communication. The better you know someone, the less 'bandwidth' you need.

(3) people's ability to carve wisdom out of data; as T.S. Eliot puts in his poem, *The Rock*:

> Where is the life we have lost in living?
> Where is the wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?

And, we may add, Where is the information we have lost in data?

(4) people informational ethics (do not distort the message; do not hoard information; make your communications with your peers easy to understand and meaningful; impart knowledge and wisdom).

I believe that these, and not the size of the electronic pipes, the bandwidth of the wires and the speed of our information appliances are the really scarce resources for us. Here's an interesting quote, from someone named Linda Ray Pratt (and I have no idea who she is) that I found the other day:

> "Philosophical habits of mind do not come quicker through fiber optics. Clear thinking is not aided by better dot resolution. Understanding ourselves and feeling for others does not come with a software upgrade."

In fact, if you think about it, most of our problems -- including privacy, access to information, miscommunications, security of communications and or failed partnerships -- arise because of some deficit in the really scarce resources.

So, let me summarize

I believe our ability to weave together well-working, successful partnerships, will depend on our ability and willingness to turn our croquet playing space into a federated environment. In this playing space, we will need to:

- learn that one size does not fit all;
- re-invent distribution partnerships when distribution costs are shrinking;
- learn to do partnerships in a world of increasing returns;
- leverage the core competencies of each partner;
- base partnerships more on cultural compatibility, trust and ethical codes, rather than exclusively contracts and legal agreements (trust, but verify, however!);

- make our partnerships a more effective market-based coordination mechanism;
- share more information, values and expectations more often with our partners;
- maintain the playing filed -- the information continuum -- in top shape at all times; and
- learn the meaning of subsidiarity in our partnership arrangements.

I have two parting thoughts, prudently, none of them mine. I thought you might, by now, have had enough of those...

- When pondering your next move, it is probably appropriate to remember a saying coined William Gibson, the inventor of *Cyberspace*: "The future is already here; it's just that it hasn't been evenly distributed".
- Finally, in respect of the issues discussed, I suggest that you will find ample guidance in the following words attributed to F. Scott Fitzgerald: "The test of a first-rate intelligence is the ability to hold two opposed ideas in mind at the same time, and still retain the ability to function. One should, for example, be able to see that things are hopeless and yet be determined to make them otherwise."

Thank you, and may our individual and collective croquet games improve!

CLOSING REMARKS

CLOSING REMARKS

G.J. Brackstone[1]

We have covered in these last two and a half days a vast and varied subject. One on the issues of content of statistical data, of technology, of methodology, and on the business side of dissemination as well. So we have covered a lot of ground, starting from Peter Hicks' vision of the kind of database that he sees needed for government policy, particularly social policy, to the discussions this morning more oriented towards the technologies on the business side and the partnerships of delivering information for users.

I think it would be impossible to summarize everything that has been covered. It would also be presumptuous of me since I was not able to attend all of the sessions. But I think there are a couple of themes that pervade what was being covered. First, clearly, is technology both in terms of the impact of computing power and how we process and manage and analyze our data, and the impact of electronic communications on how we disseminate and distribute our data. And I think the other theme, the other pressure that pervades much of what we talked about, is the financial one. The issue of doing more, delivering more with fewer resources and, as we have heard this morning, on the revenue side the impact of the extent to which statistical agencies have to cover their costs through dissemination of their data.

Before we end this formal part of the Symposium I must offer thanks, starting with our Organizing Committee. The Committee has done a lot of work for a number of months to organize this Symposium. I would like to acknowledge the committee chairman, Jean-Louis Tambay and the three other members, Georgia Roberts, John Berigan and Jean Dumais. They of course have been assisted by many people including in particular four resource persons: Josée Morel, Sophie Arsenault, Christine Larabie and Nick Budko. There are also a large number of unnamed volunteers who have helped us out in many ways and I would like to thank them too. I would like to thank very much our interpreters for two and a half days of solid work. I think that they have served us very well indeed. To all our presenters and discussants and panellists thank you also for making this Symposium possible. The people still to do their work this afternoon, the demonstrators, we thank in advance as well.

We would very much like to have comments positive or negative about this Symposium and what you think might have been done differently. I certainly received some interesting and useful comments during some of the breaks. The more of that kind of input we can receive, the more useful it would be for us. We have covered a broad range of topics. I know that not all of them would have been of interest to everyone, but I hope that everyone found something of interest and something new.

So I would like to thank all for participating, on behalf of Statistics Canada.

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26-J, R.H. Coats Bldg., Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

c. 3