



Natural Resources
Canada

Ressources naturelles
Canada



The Canadian Database of Geochemical Surveys: historical overview and current challenges

S.W. Adcock and W.A. Spirito

**Geological Survey of Canada
Current Research 2024-1**

2024

Geological Survey of Canada
Current Research 2024-1

**The Canadian Database of Geochemical Surveys:
historical overview and current challenges**

S.W. Adcock and W.A. Spirito

2024

© His Majesty the King in Right of Canada, as represented by the Minister of Natural Resources, 2024

ISSN 1701-4387

ISBN 978-0-660-69551-8

Catalogue No. M44-2024/1E-PDF

<https://doi.org/10.4095/332490>

This publication is available for free download through the NRCan Open Science and Technology Repository (<https://ostrnrcan-dostrnrcan.canada.ca/>).

This publication is also available through the Government of Canada Publications website at (<https://publications.gc.ca/>).

Recommended citation

Adcock, S.W. and Spirito, W.A., 2024. The Canadian Database of Geochemical Surveys: historical overview and current challenges; Geological Survey of Canada, Current Research 2024-1, 20 p. <https://doi.org/10.4095/332490>

Critical review

E. Boisvert

Authors

S.W. Adcock (stephenw.adcock@nrcan-rncan.gc.ca)

*Natural Resources Canada
Geological Survey of Canada
601 Booth Street
Ottawa, Ontario
K1A 0E8*

W.A. Spirito (retired)

*Natural Resources Canada
Geological Survey of Canada
601 Booth Street
Ottawa, Ontario
K1A 0E8*

Correction date:

Information contained in this publication or product may be reproduced, in part or in whole, and by any means, for personal or public non-commercial purposes, without charge or further permission, unless otherwise specified.

You are asked to:

- exercise due diligence in ensuring the accuracy of the materials reproduced;
- indicate the complete title of the materials reproduced, and the name of the author organization; and
- indicate that the reproduction is a copy of an official work that is published by Natural Resources Canada (NRCan) and that the reproduction has not been produced in affiliation with, or with the endorsement of, NRCan.

Commercial reproduction and distribution is prohibited except with written permission from NRCan. For more information, contact NRCan at copyright-droitdauteur@nrcan-rncan.gc.ca.

The Canadian Database of Geochemical Surveys: historical overview and current challenges

S.W. Adcock^{1*} and W.A. Spirito^{1,2}

Adcock, S.W. and Spirito, W.A., 2024. The Canadian Database of Geochemical Surveys: historical overview and current challenges; Geological Survey of Canada, Current Research 2024-1, 20 p. <https://doi.org/10.4095/332490>

Abstract: The Canadian Database of Geochemical Surveys (CDoGS) is a sophisticated relational database that holds a comprehensive catalogue of regional geochemical surveys carried out across Canada since the 1950s. The contents are made accessible via a public-facing Government of Canada website. The catalogue provides extensive metadata for each survey, including links to publications and digital data where possible. In addition to high-level metadata, chemical analyses for individual surveys are being added gradually. So far, 1600 surveys have been catalogued, and analytical data for 300 surveys have been incorporated, for a total of 13 000 000 individual analytical measurements.

This paper begins with a discussion of the overall IM design philosophy behind CDoGS, then moves on to give some practical examples of how the system can be used, and concludes by highlighting some of the ongoing challenges in building and maintaining the system.

Résumé : La Banque de données de levés géochimiques du Canada est une banque de données relationnelle sophistiquée qui constitue un répertoire exhaustif des levés géochimiques à l'échelle régionale effectués au Canada depuis les années 1950. Le contenu peut être consulté au moyen d'un site Web public du gouvernement du Canada. Le répertoire fournit des métadonnées détaillées pour chaque levé, y compris des liens vers des publications et des données numériques, dans la mesure du possible. En plus des métadonnées de haut niveau, les résultats d'analyses chimiques pour les levés individuels sont ajoutés progressivement. Nous avons répertorié plus de 1 600 levés à ce jour et avons intégré les données analytiques de 300 levés, pour un total de 13 millions de mesures analytiques individuelles.

L'article commence par une discussion sur la philosophie générale de la gestion de l'information qui soutient la banque de données, puis donne quelques exemples pratiques sur la façon d'utiliser le système, et conclut en soulignant certains des défis actuels liés à la création et à l'entretien du système.

¹Natural Resources Canada, Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario K1A 0E8

²Retired

*Corresponding author: S.W. Adcock (email: stephenw.adcock@nrcan-rncan.gc.ca)

INTRODUCTION

The Canadian Database of Geochemical Surveys (CDoGS) is the cornerstone of a long-running GSC activity to improve the public accessibility of Canadian geochemical data. Since the 1950s, federal and provincial geological agencies, as well as exploration companies, universities and consultancies have carried out several thousand geochemical surveys across Canada. These surveys have been completed at scales ranging from reconnaissance to very detailed and have mainly been in support of mineral exploration, but more recently for resource assessment and environmental research as well (Garrett et al., 2008). Obtaining the raw data associated with these surveys (i.e. sample locations and chemical analytical results) has commonly been a challenge. Inadequate metadata, unobtainable digital data, obsolete digital file formats, and highly variable reporting practices are just a few of the problems. The primary goal of CDoGS is to remove, or at least reduce, these difficulties.

The CDoGS is a comprehensive data management system centred on a sophisticated relational database. In order to facilitate access to the database, a website was created to present the contents of the database in a manner that strives to be user-friendly. The website URL is <https://geochem.nrcan.gc.ca>. A goal of the whole system is to adhere to the 'FAIR' (Findability, Accessibility, Interoperability and Reusability) guiding principles for scientific data management and stewardship (Wilkinson et al., 2016). Adherence poses many challenges, both technical and administrative, which are discussed extensively below.

Earlier papers (Adcock et al., 2013; Spirito et al., 2013) described some of the scientific and technical challenges involved in managing geochemical data. This paper focuses more on the overall Information Management (IM) challenges in maintaining CDoGS. Given its IM focus, the paper is unavoidably burdened with a large number of acronyms. As an aid to the reader, definitions for each acronym are given in an appendix.

PLACING CDOGS IN CONTEXT

The CDoGS resides within an extraordinarily complex IM/IT ecosystem that is constantly evolving. In order to gain a reasonable understanding of its architecture, it is necessary to have a certain degree of understanding of many inter-related IM/IT subjects, both technical and administrative. CDoGS needs to be placed within a wider context of government scientific data. This wider context can be examined from a variety of perspectives, as outlined below:

- the Canadian geoscience context,
- the federal geospatial/geoscience context,
- the federal IM context,

- the Global IM context, and
- the Global geoscience context.

THE CANADIAN GEOSCIENCE CONTEXT

Canada functions under a system of federal government, whereby responsibilities are divided between the federal, provincial and territorial governments. Provinces have the primary responsibility for the control of natural resources, which has led all of them, except Prince Edward Island, to establish their own Geological Surveys. However, the federal government in Ottawa has always played a major role in mineral development across the country. The Geological Survey of Canada (GSC) was established in 1842, 25 years before Canada itself achieved nationhood, and over a century before Newfoundland became the final province of the Confederation. The GSC has always worked very closely with its provincial and territorial counterparts. Many of the largest and most significant geochemical surveys have been joint federal-provincial activities, in support of mineral development. This shared responsibility does lead to complications in the management of the data generated from the surveys that have been exacerbated in recent years by two factors: a) the significant downsizing that has occurred over the past 25 years at the GSC and all of its provincial counterparts, and b) the need to adhere to multiple standards and protocols for making the data available digitally over the Internet.

The Canadian Geoscience Knowledge Network (CGKN), established in 1998, was an ambitious initiative to provide a unified WWW portal to access the data holdings of the GSC and its provincial counterparts (Grunsky and Broome, 2001; Rupert et al., 2002; Grunsky et al., 2003). Active development after 2003 was minimal, and despite some successes the effort eventually collapsed as funding sources dried up. The *cgkn.net* domain name was relinquished a few years later. In hindsight, it seems that the initiative was overly ambitious. The underlying architecture required the establishment of a distributed network of Z39.50 servers, each providing metadata that conformed to the U.S.-based FGDC standard (National Information Standards Organization, 2003; Federal Geographic Data Committee, 2000). The technology was difficult to implement and beyond the resources of most of the participants. The CDoGS Z39.50 Unix server was shut down when CGKN stopped being maintained.

THE FEDERAL GEOSPATIAL/GEOSCIENCE CONTEXT

After the demise of CGKN, the various federal, provincial and territorial agencies each implemented their own WWW interfaces to their data holdings with very little consideration

for interoperability or a ‘common look and feel’ (CLF). At the GSC, CGKN was succeeded by the Geoscience Data Repository (GDR) (Grunsky et al., 2003). The GDR never evolved beyond a home page that presented a set of links to independently managed data sets. The GDR website itself is no longer maintained, but numerous snapshots of it can be viewed via the Wayback Machine (e.g. https://web.archive.org/web/20080621184245/http://gdr.nrcan.gc.ca/index_e.php captures the site as it existed on June 21st, 2008). The website became inactive in 2012, but some of the independent data sets continue to refer to the GDR acronym.

Recent efforts at the GSC have focused on complying with larger federal government initiatives to provide unified access to digital data. The ‘Open Government Initiative’ was launched in 2011 with three streams: Open Information, Open Data, and Open Dialogue. Within this context, the GSC Strategic Plan for 2013–2018 identified ‘Open Geoscience’ as one of its five major priorities (Geological Survey of Canada, 2014). The most recent strategic plan re-affirms this priority within the broader priority of ‘Geoscience for society’ (Geological Survey of Canada, 2018).

As part of the dramatic downsizing that occurred across the Federal government in the mid-1990s, the GSC was merged with Surveys, Mapping and Remote Sensing Sector (SMRSS) within Natural Resources Canada (NRCan), to become the Earth Sciences Sector (ESS). This merger had several consequences that had a dramatic impact on the evolution of the GSC’s WWW data-delivery strategy. At the time of the merger, SMRSS was already engaged in several major projects to improve WWW data delivery. Most of these were being funded by the GeoConnections program, and much of the data was made publicly accessible via the Canadian Geospatial Data Infrastructure (CGDI) website (at <https://cgdi.ca/> until 2018 and now at <https://natural-resources.canada.ca/science-and-data/science-and-research/geomatics/canadas-spatial-data-infrastructure/10783>).

The GeoGratis FTP site was a significant component of CGDI. As the GSC and SMRSS became more tightly integrated there was an emphasis on developing integrated websites. This led eventually to a greatly enhanced HTML version of GeoGratis, capable of spatially querying the GEOSCAN bibliographic database. GEOSCAN, like CGKN, began as a cooperative project between the GSC and provincial surveys, with the goal of providing a single catalogue of all the publications of the GSC and its provincial counterparts across Canada (Blair et al., 1993; Kopf-Johnson, 1994; Blair, 2001). Downsizing in the 1990s forced the GSC to step back from its commitment to catalogue publications of the provincial surveys. But the integration with SMRSS led to its scope expanding to include all of ESS.

The enhanced GeoGratis map query tool had a very short life, as it was overtaken by the Federal Geospatial Platform (FGP) in 2017 — an initiative with a very similar goal, but encompassing the whole of the Federal Government, not just ESS. As with the GDR, the history of the GeoGratis website

can be examined via the Wayback Machine, using the URL <http://geogratis.gc.ca/>. Note that the technology used to display the GeoGratis web pages leads to problems when viewing some of them via the Wayback Machine.

The FGP is accessible via two portals: public and internal to the Federal government (<https://maps.canada.ca/en/index.html>). The initiative began to take shape in 2012 with the creation of the Federal Committee on Geomatics and Earth Observations (FCGEO) (Shukle, 2014; Loubier, 2015), but progress has been slow.

THE FEDERAL IM CONTEXT

The FGP is just one of numerous initiatives that have been launched over the past 20 years that is designed to provide a more consistent and open approach to the delivery of digital data and information across all Federal government departments (Clarke, 2019). Treasury Board Secretariat (TBS) is typically the lead agency for these initiatives, which include:

- Common Look and Feel (CLF). An effort to ensure that all Federal government websites have a similar appearance and core functionality. Staff within TBS, led by Paul Jackson, created the Web Experience Toolkit (WET), as a way of achieving conformance to the CLF guidelines.
- Web Content Accessibility Guidelines (WCAG). These guidelines are designed to ensure that websites are accessible to people with disabilities. The Jodhan court case (Jodhan v. Canada, 2010) led to the requirement that all public-facing Federal government websites must be WCAG-compliant. An unfortunate consequence of this decision was that many web pages were taken offline because of technological problems in achieving compliance, and they were never redesigned because of insufficient resources.
- Government of Canada Core Subject Thesaurus, accessible at <https://open.canada.ca/data/en/dataset/d4a0e406-eea9-41a7-bcae-28c31f3b9c65> (Renaud, 2004; Treasury Board of Canada Secretariat, 2013). This thesaurus is intended to cover all of the fields treated in information resources of the Government of Canada, but it is relatively small, consisting of only about 5,000 terms. Because of the great variety of subjects covered by the thesaurus, its terminology is rather general. By design, it does not include specialized terminology used in specific and limited disciplines. For example, geochemistry is a term within the thesaurus, but there are no terms that are more detailed.

The difficulties of complying with CLF and WCAG led to many of the GSC’s web pages being taken offline. They remain accessible via the Wayback machine, using the URL <http://nrcan.gc.ca>.

THE GLOBAL IM CONTEXT

The first efforts at using relational database technology to manage geochemical data at the GSC were made by Steve Adcock in 1988, using Oracle software on a VAX mini-computer. Information management advances since then have been extraordinary. Digital data dissemination methods at the GSC evolved from mainframe 9-track tapes in the early 1980s, through diskettes and CDs, to a near-total reliance on the Internet. Oracle was the only purveyor of relational database management system (RDBMS) software in the mid-1980s. Nowadays, the GSC employs a variety of software, including Oracle, PostgreSQL, Microsoft SQL Server, and Microsoft Access®.

In the late 1980s, GIS software was still in its infancy. Almost all of the maps at the GSC were still produced manually, using traditional drafting techniques. Advances in computing power spurred most of the progress in GIS, but there were two additional factors. One was the creation of the GPS satellite system. The second was free-and-easy access to high-resolution satellite imagery, epitomized by the appearance of Google Earth™ in 2005 (and earlier incarnations created by Keyhole Inc.).

The rise of microcomputers in the 1980s triggered a revolution in software development. The cost of writing software plummeted and the customer base expanded rapidly. Until the mid-1980s, the GSC released its digital geochemical data as simple text files that were designed to be read by FORTRAN software. Moving forward, management of geochemical data devolved from a small group of FORTRAN programmers to individual scientists. Scientists were free to manage the data in whichever way they chose. The end-result was a collection of incompatible data structures, stored in an endless variety of proprietary file formats. The situation gradually became simpler, as Microsoft® Excel® emerged as a de-facto standard for storing and transferring geochemical data. But MS Excel® is not designed to be a data management tool, and reliance upon it has created many challenges, that will be discussed in detail in later sections on IM standards and data integrity.

THE GLOBAL GEOSCIENCE CONTEXT

Government scientific organizations enjoyed a long period of growth across the developed world in the years following the Second World War. Growth began to slow down in the 1970s, and many organizations began shrinking in the 1980s. The GSC has followed this global rise and fall. The very large regional geochemical surveys that were conducted by GSC scientists have now ceased. Modern surveys are far fewer and much smaller. This pattern is worldwide (Agnew, 2017). However, many of the samples that were

collected have been safely archived, and are available for re-analysis by newer and greatly improved analytical techniques. Although the total number of samples collected is growing very slowly, the number of analytical values is growing rapidly.

CONTEXTUAL SUMMARY

The various contexts discussed above have led directly to some of the fundamental characteristics of CDoGS:

1. **Inclusivity** — the system catalogues all geochemical surveys across Canada, not just those led by the GSC. It is impractical to rigidly divide the responsibility for management of geochemical samples (storage, re-analysis, publication, etc.) between the GSC and its provincial counterparts.
2. **Software Independence** — the system seeks to minimize any dependencies on proprietary software and data formats. Technology continues to evolve extremely rapidly, and the system must be agile enough to switch to new strategies and paradigms.
3. **Standards Adherence** — IM standards are followed wherever they exist. This allows easy conformance with various initiatives such as CLF, WCAG, and FGP to facilitate interoperability.
4. **Simplicity** — complex software architectures are avoided. For example, the website relies on static HTML pages, not dynamic pages linked to a database. Custom in-house software is time-consuming to create and maintain, and should be kept to a minimum.

INFORMATION MANAGEMENT STANDARDS, OPEN SOURCE SOFTWARE

The CDoGS strives to adhere to the ‘FAIR’ principles for scientific data management. These principles deal with findability, accessibility, interoperability, and reusability of scientific data and its stewardship (Wilkinson et al., 2016). A fifth principle of ‘Sustainability’ is also critically important; it is described in more detail in a later section. Adherence to all of these principles is greatly improved by conforming to internationally recognized IM standards. The relevant standards adhered to by CDoGS include the following:

Structured query language (SQL)

Structured query language (SQL) lies at the core of manipulating data within relational databases. The standard specification for SQL has gone through nine iterations between 1986 and 2016. Commercial vendors of relational

database management system (RDBMS) software generally support a subset of the official standard, and offer various enhancements. Additionally, SQL is a declarative language, not a procedural one. The various vendors all offer extensions to SQL to deliver a procedural programming capability (PL/SQL in the case of Oracle, T-SQL in the case of SQL Server). Maintenance of the CDoGS database relies heavily on SQL scripts. The scripts have been written using a subset of the official SQL standard, and they do not rely on any procedural language extensions. This greatly increases the portability of the database between different RDBMS platforms, at the expense of lengthier scripts. Portability has been tested with SQL Server, Oracle, Ingres and MS Access[®].

The CDoGS database currently uses SQL Server as its RDBMS software, but it would be easy to switch to other software.

Extensible markup language (XML)

Data manipulation within the CDoGS relational database is accomplished via SQL, but data manipulation outside the database is accomplished primarily by the transformation of XML documents. This includes both the loading and exporting of data. The CDoGS website presents the database contents primarily as XHTML5 web pages, KML maps, and MS Excel[®] spreadsheets. The structure of the XML files is tightly constrained by XSDs, and transformation is performed using XSLT. XML/XSD/XSLT are a powerful set of technologies for data transformation. XSLT, like SQL, is a ‘declarative’ programming language, in contrast to procedural languages such as C and FORTRAN. Within the context of data transformation, it leads to simpler programs.

International Organization for Standardization (ISO) 19115

Adherence to geospatial metadata standards has been a goal and a challenge at the GSC since the 1990s. The CSDGM standard developed by the FGDC in the 1990s was superseded by the ISO 19115 standard. The ISO 19115 is an abstract specification. The ISO 19139 is an XML implementation of the specification. It is possible to implement profiles within the ISO 19115 framework to better address the requirements of a particular organization, whilst remaining fully compliant with ISO 19115.

The FGP does precisely this by requiring metadata to conform to the Harmonized North American Profile (HNAP) (Moellering et al., 2008; Natural Resources Canada, 2015). All of the data that are needed to build HNAP-compliant metadata records are contained within the CDoGS database.

Open Geospatial Consortium (OGC) standards

The Open Geospatial Consortium (OGC) has established numerous standards, in pursuit of the ‘FAIR’ principles. The CDoGS website makes extensive use of the keyhole markup language (KML) standard for delivering geospatial data. Web mapping services (WMS) is another OGC standard that is used within the CDoGS system to deliver map images; it has been a critical component of recent NRCAN and Federal Government initiatives, including the enhanced version of GeoGratis and FGP. The CDoGS WMS is accessible at: https://geochem.nrcan.gc.ca/cgi-bin/GSC_Geochemistry/wms?service=WMS&request=GetCapabilities&

Open source software

The CDoGS has been developed under the MS Windows[®] operating system (OS). Some aspects of the CDoGS system are OS-dependent, but there has been a deliberate effort to isolate and minimize these dependencies, such that migration to an alternative OS such as Linux would not be difficult. Reliance on open source software plays a major role in minimizing dependencies. Important software packages include:

- Saxon XSLT processor (<http://www.saxonica.com/welcome/welcome.xml>)
- MapServer (<https://mapserver.org/>)
- QGIS (<https://www.qgis.org>)

FUTURE DIRECTIONS

The architecture of CDoGS will continue to evolve as new data management paradigms emerge. The precise details of the evolution will be critically dependent on the overall evolution of scientific IM, which remains unclear, both within the GSC and in a global context.

Findability

‘Findability’ is a huge issue. Expressed simplistically, findability is centred on the provision of keywords, which can be used by Internet search engines. In the early days of the WWW, keywords were specified explicitly in the <head> element of HTML pages, by using the <meta> tag. But this mechanism was abused by people attempting to improve their website’s search engine ranking, and modern search engines de-emphasize the content of <meta> tags. Recent techniques to enhance findability are more sophisticated, and considerably harder to understand and implement.

The ‘Semantic Web’ has the potential to allow data to be integrated across the Internet, which will greatly simplify ‘findability.’ Broadly speaking, it is a re-visioning of

the WWW that will allow data to be shared easily across websites (and therefore ‘found’) instead of each application holding on to the data for itself. The Semantic Web is a concept, rather than a specific project (World Wide Web Consortium, 2009).

The Semantic Web vision is highly dependent on ‘ontologies.’ As explained by World Wide Web Consortium (W3C):

Ontologies define the concepts and relationships used to describe and represent an area of knowledge. Ontologies are used to classify the terms used in a particular application, characterize possible relationships, and define possible constraints on using those relationships. In practice, ontologies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only).

An ontology that captures all of the complexities of geochemical data still needs to be developed.

Two initiatives to enhance findability that are being coordinated by the W3C stand out as pointing the way forward: Schema.org and DCAT.

Schema.org

Quoting from Wikipedia (Wikipedia contributors, 2019):

Schema.org is a collaborative community activity with a mission to “create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.” Webmasters use this shared vocabulary to structure metadata on their websites and to help search engines understand the published content, a technique known as search engine optimization.

The full hierarchy of the Schema.org vocabulary is accessible at <https://schema.org/docs/full.html>. Perusing this hierarchy, it is clear that in its current state of development it does not address the needs of scientific data sets. If one attempted to fit a geochemical survey into the hierarchy of schema.org, it would be very abstract and look something like:

- Thing
 - Intangible
 - Structured Value

This vagueness arises because there are no suitable, specific terms to describe a geochemical survey and it can only be described in very abstract terms. However, there are hundreds of specific terms in the hierarchy (e.g. GlutenFreeDiet, FlightReservation) that allow many things to be described precisely. For example, an online restaurant menu could tag individual items with ‘GlutenFreeDiet.’

In the absence of detailed guidance for geochemical surveys and other scientific concepts, scientists would inevitably describe surveys inconsistently. The ‘shared vocabulary’ needed to consistently describe geochemical surveys does not yet exist.

Data catalog vocabulary (DCAT)

Quoting again from Wikipedia (Wikipedia contributors, 2018-08-31):

Data Catalog Vocabulary (DCAT) is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. By using DCAT to describe datasets in catalogs, publishers increase discoverability and enable applications to consume metadata from multiple catalogs. It enables decentralized publishing of catalogs and facilitates federated dataset search across catalogs. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

The objectives of DCAT are essentially the same as Schema.org, and the two initiatives strive to be compatible with one another. The Schema.org methodology works by embedding HTML5 tags within HTML pages. DCAT information can be presented in many forms; one way is to embed it in HTML pages as HTML-RDFa. As with Schema.org, the absence of a ‘shared vocabulary’ is a major impediment to DCAT’s usefulness.

Geoscience Markup Language (GeoSciML) is a major effort to provide a ‘shared vocabulary’ for geology. Version 4.1 was released in 2017 (Boisvert et al., 2017). The GSC has played a major role in the development of GeoSciML, through the involvement of Eric Boisvert, Boyan Brodaric, and François Létourneau along with an international team. Within the overall GeoSciML model, there is a ‘Laboratory and Analysis’ package, which provides a solid framework for capturing metadata about geochemical analyses. But the vocabulary is not sufficiently detailed to meet the needs for fully describing geochemical data.

What would a comprehensive ‘shared vocabulary’ for geochemical surveys look like? The two most important aspects that are still undeveloped centre on a) description of the sample medium and b) description of the laboratory procedures. Understanding exactly how a sample was collected, processed, and analyzed is critically important to data interpretation. Sample media vary enormously, from trout livers (Warren et al., 1971) to lunar rock samples. Similarly, collection procedures are endlessly variable; a bedrock sample, for example, may be collected from a surface outcrop using a hammer, or from underground by any one of numerous drilling techniques. Laboratory procedures are similarly extremely variable; furthermore, they are evolving rapidly. Instruments such as portable XRF blur the distinction between field and laboratory. A straightforward unstructured

list of keywords covering all sample media, collection procedures, and analytical methods would be, in itself, a significant undertaking. But to maximize its usefulness, the list should have some structure, and also some authority.

Authority typically derives from endorsement by a standards-setting group such as ISO, W3C or OGC. Authority typically implies a ‘controlled vocabulary’, where there is some sort of governance and rules to follow if the list needs to be changed. Structure can take on different forms. It may be a simple hierarchy (aka taxonomy), a more sophisticated thesaurus, or a complex ontology. GeoRef (Goodman, 2008) is an example of an extensive thesaurus that is widely used by libraries for cataloguing geoscience publications. ISO 19115, DCAT, and GeoSciML are all examples of ontologies.

As previously discussed, ontologies can be very complex. Geography markup language (GML) is a good example of complexity. The PDF version of GML v3.2.2 is over 400 pages long (Portele, 2016). Within the context of the WWW, ontologies are defined using OWL (web ontology language), a W3C standard. There have been efforts to extend the standards into specific domains for both GML and ISO 19115. Groundwater markup language (GWML) is an extension of GeoSciML to cover groundwater (Boisvert and Brodaric, 2012; Brodaric et al., 2018). Similarly, there is a ‘Biological Profile’, which extends ISO 19115 (Mize, 2012). Any authoritative attempt to create an ontology for geochemical surveys would be a major multi-year undertaking involving many people and organizations.

Digital Object Identifiers (DOIs) have the potential to greatly simplify the discovery of data on the WWW. To date, DOIs have been used primarily to simplify access to published documents, especially scientific papers. But they have the potential to be applied much more widely, to any ‘digital object’, including items such as geochemical surveys which exist only as abstract entities. A unique DOI could be assigned to each survey in the CDoGS catalogue. DOIs could even be assigned to individual samples. The object’s DOI can then be used to refer to it unambiguously from any digital resource. Quoting from Paskin (2010):

The DOI system provides identifiers which are persistent, unique, resolvable, and interoperable and so useful for management of content on digital networks in automated and controlled ways.

Juty et al. (2020) expand on the importance of identifiers in the context of the FAIR principles.

Sustainability

The explosive growth of data and information on the WWW over the past 25 years raises many questions about sustainability. A website which fully addresses the FAIR principles is of little use if the server on which it is running is shut down. On the other hand, many websites continue to exist long after they have ceased to be actively maintained.

In so doing, they frequently propagate inaccurate or false information. Sustaining a complex information delivery system such as CDoGS requires a major commitment. But long-term commitments are hard to maintain — individuals move on to other challenges, and the organization’s priorities and resources change. A scientific or academic organization will typically have a relatively small number of mission-critical systems that must be maintained, plus a much larger number of ‘optional extras’. It is hard to identify any of the GSC’s scientific data delivery systems as being truly mission-critical. In this environment, it is hard to guarantee their long-term futures.

As the WWW evolves and the first wave of content-generating researchers retires, sustainability is likely to become a major issue. Artificial Intelligence (AI) may evolve quickly enough to be at least a part of the solution, but as yet there is no clear overall solution.

Much of the modern IM/WWW ecosystem is dependent on public generosity and volunteers. Wikipedia is a prime example of this funding model, but there are many others. Prominent examples used by geochemists include R, QGIS and OpenLayers. The long-term viability of open data and software remains unclear. In such an uncertain environment, digital data custodians need to put mechanisms in place to ensure that the data are preserved in ‘worst-case’ scenarios.

The GSC is fortunate in having an Open File publication series, which facilitates the safe long-term archiving of large, complex, digital entities such as databases.

CDOGS – A BRIEF OVERVIEW OF ITS PUBLIC INTERFACE

The CDoGS website has two principal components. The first is a catalogue of geochemical surveys carried out across the country, complete with rich, extensive metadata as well as links to additional resources. The second is a data warehouse of analytical data associated with a subset of the catalogued surveys. The catalogue currently comprises about 1600 surveys. The data warehouse contains analytical data for almost 300 of approximately 500 that are deemed to be of strategic, long-term value.

The HTML pages on the website are viewable in any modern web browser, on any kind of device, but many of the pages are very large and contain big tables. Therefore, they are best viewed on a desktop computer with a lot of memory and a large display, via a fast Internet connection. The website relies extensively on KML files for displaying geospatial data. The KML data can be viewed by many software applications, but the website KML files are optimized for viewing using the desktop version of Google Earth™ (Spirito and Adcock, 2010).

Examples of querying the website

The website catalogue can be searched in a variety of ways, depending on the end user's goal. All of the major entities within the catalogue (surveys, publications, projects, etc.) are accessible via simple HTML index tables that can be filtered and sorted. The user can also search for surveys geospatially, either via a map query interface or by KML index maps. Surveys can have very complicated histories that may be reported in more than one publication, so the website is structured in a way to provide users with multiple access points to the information. This is illustrated by the following use cases where the end user:

1. is interested in the geochemistry of vanadium across Canada,
2. is interested in all of the NGR lake sediment data across Canada,
3. has a vague memory of a survey carried out by Don Hornbrook in the early 1970s,
4. is interested in all of the geochemical data in and around the Athabasca Basin, and
5. wants to obtain the geochemical data that were published in GSC Open File 1335.

1. Vanadium

There is no easy way to determine which of the 1600 catalogued surveys include data for vanadium (V). Searching through all of them would involve a very laborious reading of all of the publications, which would be immensely time-consuming. However, it is trivial to identify which of the 300 surveys in the data warehouse include vanadium data. The website includes a periodic table interface at: https://geochem.nrcan.gc.ca/pertable/content/pertable/main_e.htm, accessible from the CDoGS home page by clicking on the 'Periodic Table' link (Fig. 1). Clicking on V in the periodic

table will download a KML file showing which of the 300 surveys include V data. If you view the KML file in Desktop Google Earth™, you can use the legend in the 'Places' windowpane to get a detailed breakdown of the surveys by sample type (e.g. till, NGR stream sediment/water) or survey type (e.g. indicator minerals, resource assessment). These can be toggled on/off. Clicking on any of the stars on the map will launch a pop-up window summarizing the V data for a survey, with links to additional metadata as well as to simple KML files of the analytical data (Fig. 2).

All of the vanadium data for Canada can be downloaded in either spreadsheet (MS Excel®) or database (MS Access®) formats. These data can be accessed via the list of analyzed quantities at: https://geochem.nrcan.gc.ca/cdogs/content/tables/list_qty_en.htm. The HTML table can be filtered by typing vanadium in the 'Name' column. As of July 2023, the database contains 189 496 analytical values for vanadium. Clicking on '189 496' leads to a summary page of the data presented as a simple HTML table. This table provides a breakdown of the sample types that were analyzed for V and by what method. The summary page can also be viewed as a tree. Links are provided to the MS Excel® and MS Access® files.

2. National Geochemical Reconnaissance (NGR) lake sediment data

The National Geochemical Reconnaissance (NGR) data for stream and lake sediments are by far the largest and most important data sets within the CDoGS system (McCurdy et al., 2014). Much of the NGR data are included in the 300 surveys, and loading the remaining data are a high priority. The easiest way to identify all of the NGR lake sediment surveys is via the KML index map at: https://geochem.nrcan.gc.ca/kml/data/index/surveys_ww_e.kmz, by clicking on the 'All Surveys' link on the left hand side of the CDoGS home page. Using the legend on the left and selecting 'NGR lake (107)',

Periodic Table Index

Element List CDoGS

H																			He
Li	Be										B	C	N	O	F				Ne
Na	Mg										Al	Si	P	S	Cl				Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br			Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			Xe
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At			Rn
Fr	Ra	Ac	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts			Og
Lanthanides																			
La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu					
Actinides																			
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr					

Legend: Data available ✕ No data available

Figure 1. Periodic Table web page showing elements for which analytical data have been loaded into the database (https://geochem.nrcan.gc.ca/pertable/content/pertable/main_e.htm)

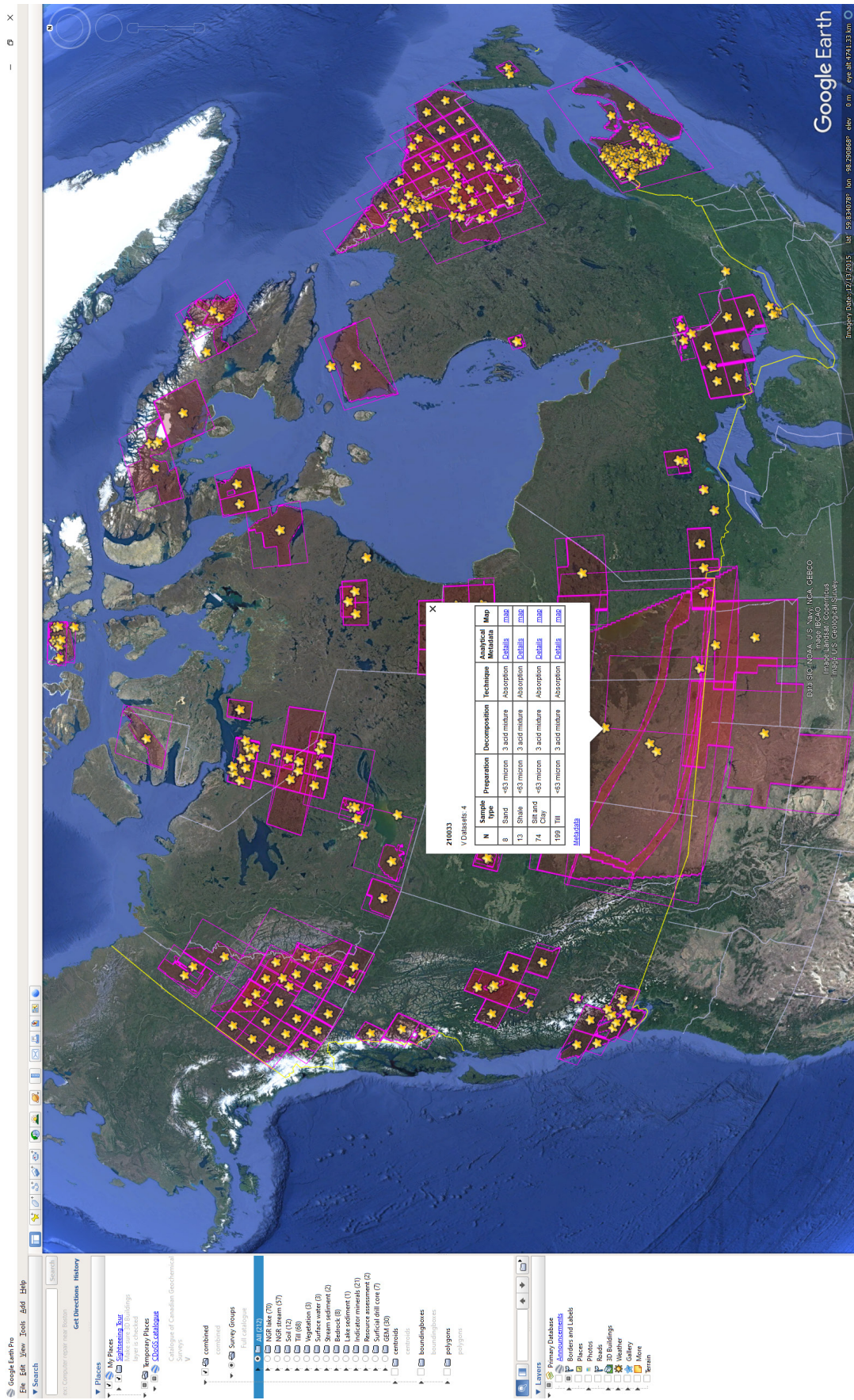


Figure 2. Index map of surveys for which analytical data for vanadium have been loaded into the database (https://geochem.nrcan.gc.ca/kml/data/index/elements/idx_elem_v_wv_e.kmz)

the 107 individual surveys can be viewed. Clicking on a star will lead to a metadata page about the survey, including links to the raw data as they were released in associated publications (Fig. 3).

From the left hand side of the CDoGS home page, clicking on the link for 'Raw Data Surveys' (https://geochem.nrcan.gc.ca/kml/data/index/surveys_ext_ww_e.kmz) will download a KML map showing the 292 surveys for which raw data have been loaded into the data warehouse. The map can then be filtered to show the 94 NGR lake sediment surveys that have been loaded. For each of these 94 surveys, the analytical data can be downloaded in a standardized format, by following links from the metadata page for each survey. The standardized data format is described in more detail below.

The user can also see a list of all of the NGR surveys by visiting the Surveys summary page at https://geochem.nrcan.gc.ca/cdogs/content/tables/_list_svy_en.htm, by clicking on the 'Surveys' Index Table link on the CDoGS home page. The list of surveys can be filtered by specifying *ngr lake* under the *Grouping* column. The list of surveys can be further filtered or sorted using the other columns, such as *Province* or *Year*.

3. Hornbrook Surveys

Don Hornbrook was the principal investigator for many geochemical surveys during his career at the GSC. The website provides two ways to examine his work. The first way is to search the Publications list at https://geochem.nrcan.gc.ca/cdogs/content/tables/_list_pub_en.htm, by clicking on the 'Publications' Index Table link on the home page. The Publications HTML table is very large. Specifying *Hornbrook* in the *Recommended Citation* column leads to a list of all of the catalogued publications where he was the author or co-author. Sorting these publications by year, they span the range 1967 to 2009. This can help to narrow the search if the user is looking for surveys carried out in a particular timeframe. Perusing the titles may lead the user to the Hornbrook activity of interest. Following the hyperlink in the *ID* column of the row of the publication identified by the user leads to a page of metadata about the publication, including links to the surveys that are connected to the publication.

The second way is to search the Projects list at https://geochem.nrcan.gc.ca/cdogs/content/tables/_list_prj_en.htm, by clicking on the 'Projects' Index Table link on the home page. Filter the table using *hornbrook* in the *Leader* column, and sort the rows using the *Date* column. Perusing this list may lead the user to the project of interest. As with the Publications table, following the hyperlink in the table row (*Key* column) leads to a page of metadata about the project, including links to the surveys that are connected to the project. For example, if the project of interest is the *Winter Works Program, Timmins - Val d'Or, Ontario and Quebec, 1971–1972*, the user can follow the various hyperlinks

on the Project Metadata page to download the actual data and view simple KML maps. This is possible because the WinterWorks project included a large lake sediment survey for which the data have been loaded into the data warehouse. Figure 4 shows the copper data for 2692 samples collected in this survey.

4. Athabasca Basin

This geographic location query for the Athabasca Basin lends itself to a geospatial search. The KML Index Maps of the surveys, accessible from the side menu on the CDoGS home page, give a quick overview of the surveys catalogued. A more powerful search involves using the map query tool ('Map Query' link on the home page) at https://geochem.nrcan.gc.ca/indexmap/content/mapserver/main_e.phtml. The tool uses standard operations to allow the user to zoom and pan to the Athabasca Basin. The initial view of the surveys is complicated because of overlapping survey extents. The surveys can be filtered using the various fields to the left of the map window. For example, filtering on *Survey Type* equal to *NGR Lake* displays a map of NGR lake surveys across Canada. The user can then zoom in to the Athabasca Basin (Fig. 5).

The map query tool is based on the MapServer software that was originally developed at the University of Minnesota (<https://mapserver.org/>). Future plans for the website include a major overhaul of the user interface.

5. GSC Open File 1335

To query for this publication, begin with the Publications list at https://geochem.nrcan.gc.ca/cdogs/content/tables/_list_pub_en.htm, by clicking on the 'Publications' Index Table on the CDoGS home page. Filter for 1335 in the *GSC Open File* column, and follow the link to the metadata page for the publication. From the publication metadata page, follow the link to the metadata page for the associated survey (till samples, Baker Lake area, Nunavut, 1975–1976). From this page, the user can follow links to obtain the raw data either in the original published format, or in the standardised format used by CDoGS. This standardised data format is described in the following section. Figure 6 is a map of Zn data for this survey.

Standardised data format

Presenting geochemical data in a standardized format when the data span an open-ended range of sample media and analytical methods is challenging. A critical simplifying assumption in CDoGS is that a sample's location can be specified by a single X-Y point. The system does not attempt to standardize any field observations that are associated with the sample because there is commonly very little consistency between geochemical surveys. For example,

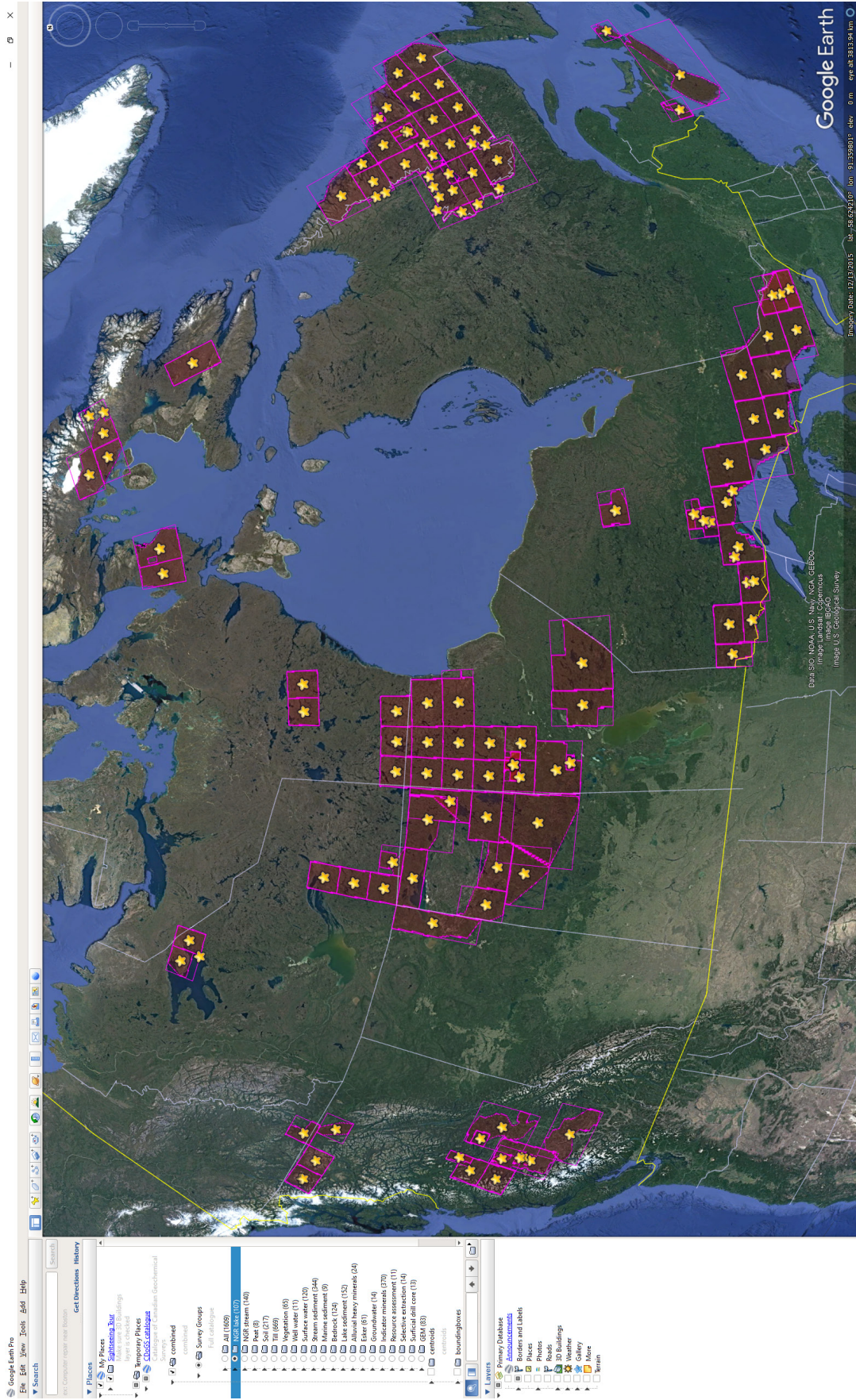


Figure 3. Index map of NGR lake sediment surveys (https://geochem.nrcan.gc.ca/km/data/index/surveys_w_w_e.kmz)

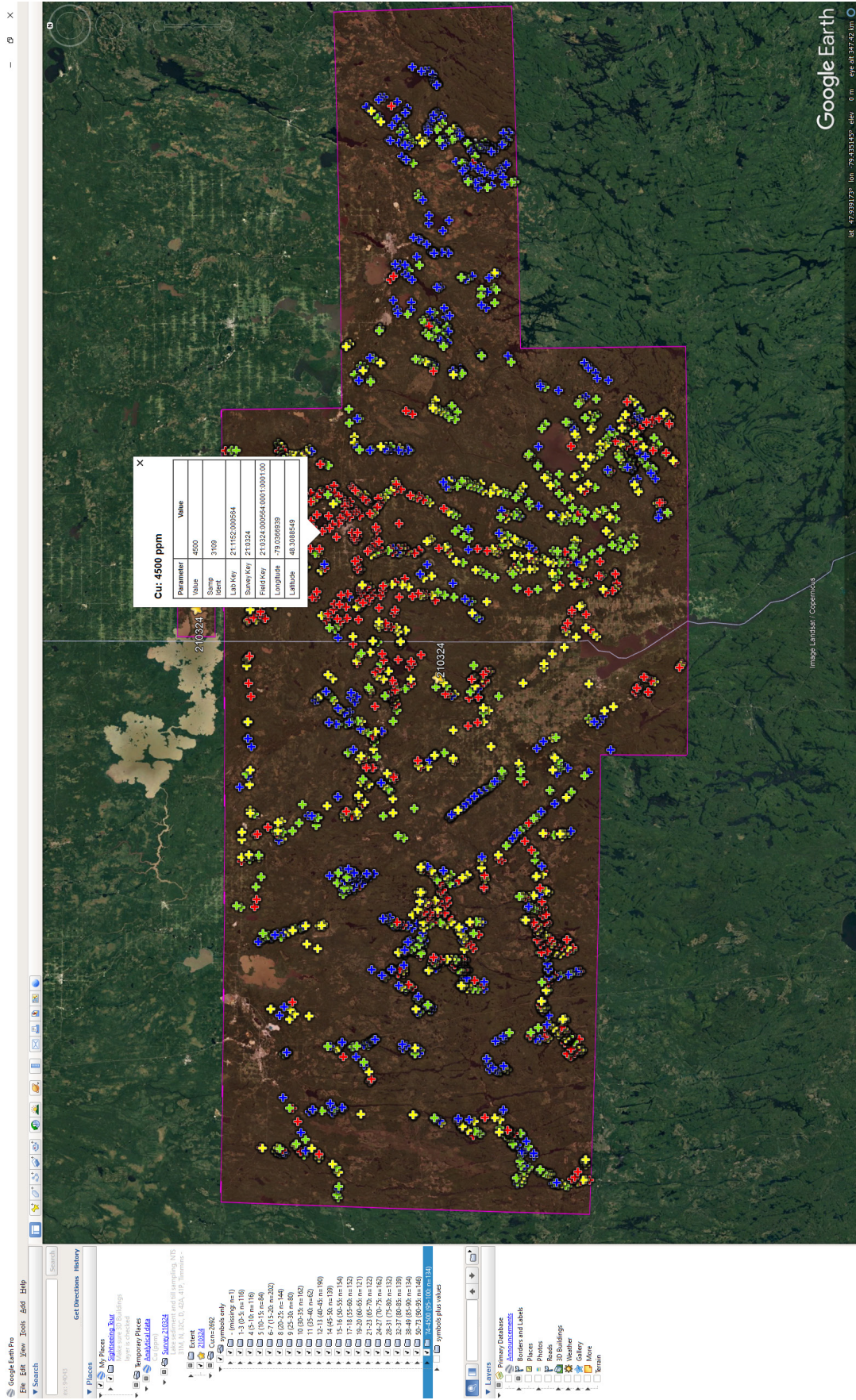


Figure 4. Map of copper in lake sediments (AAS), Winter Works Program, Timmins - Val d'Or, 1971-1972 (https://geochem.nrcan.gc.ca/kml/data/analyses/bdlj211152059850100023001_ww_e.kmz)

Index Map Query

Map Results CDoGS

Survey Count: 107

Province: All

Organisation: All

Survey Type: NGR lake

Year: All

Title:

Abstract:

Location

Longitude:


Latitude:

or


NTS:

Raw data surveys only


Limit Selection Reset

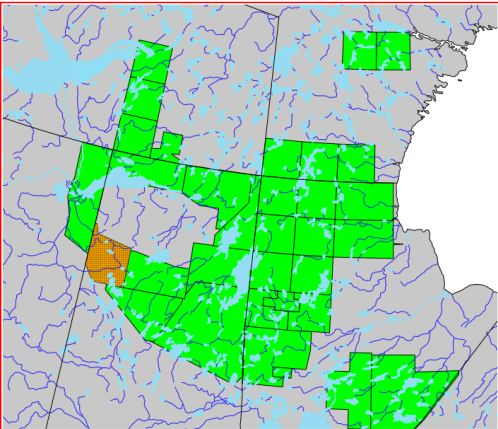


Mouse Navigation



Keyboard Navigation





NTS

Default

250K

50K

Map Limits

N 64.73838

E -88.58631

W -118.21787

S 52.17354

Scale

1: 5,757,632

km

Redraw

Submit

High Res

Figure 5. Index map query, showing NGR lake sediment surveys near the Athabasca Basin (https://geochem.nrcan.gc.ca/indexmap/content/mapservers/main_e.phtml)

Query Results

Highlighted Survey Count 1

ID	Grouping	Organisation	Year	Survey Description
210134	NGR lake	GSC-Northern Div.	1978	NGR lake sediment and water survey, NTS 74C, F, northwestern Saskatchewan, 1978

one survey may record sediment colour using the Munsell chart but another survey may use an arbitrary scheme. Field observations can be stored in the database, but currently they can only be retrieved by custom SQL queries carried out by someone with SQL expertise and knowledge of the CDoGS data model.

The data format for spreadsheets of downloadable data has three variants, depending on how the end user wishes to handle data below detection:

- numerical values, with data below detection represented as a negative value (i.e. <2 becomes -2),
- numerical values, with data below detection represented as half the detection limit (i.e. <2 becomes 1), and
- text values (i.e. <2 stays as <2).

The standardized data format includes a fixed set of sample identifier columns, followed by a well defined set of chemical analysis columns. The set of chemical analysis columns corresponds exactly to the set of columns in the associated 'Analytical Package'. The concept of an analytical package requires a basic understanding of the operational procedures of analytical laboratories. A typical laboratory

will contain a range of instruments, each of which can measure a set of different analytical quantities (Cu, Pb, pH, SO₄²⁻, etc.). Customers will send a 'bundle' of samples to the laboratory and request a 'package' of analyses from one or more of these different instruments. Within the CDoGS terminology, a particular instrument will measure a 'suite' of analytical quantities. These suites are then combined to produce a package. Each suite will contain one or more different analytical methods corresponding to different analyzed quantities. For example, an INAA suite will typically contain data for about 35 different elements. All of the methods within a suite will have the same analytical technique (e.g. ICP-MS, AAS, INAA, etc.) and the same sample decomposition (e.g. aqua regia, borate fusion, etc.).

The CDoGS data model requires that an analytical package is tied to a specific laboratory, but suites and methods may be shared by different laboratories. The package/suite/method hierarchy leads to a relatively small number of distinct packages (currently about 400). A package may include several different analytical techniques (e.g. ICP-MS and INAA), but a suite will always correspond to just one specific technique.

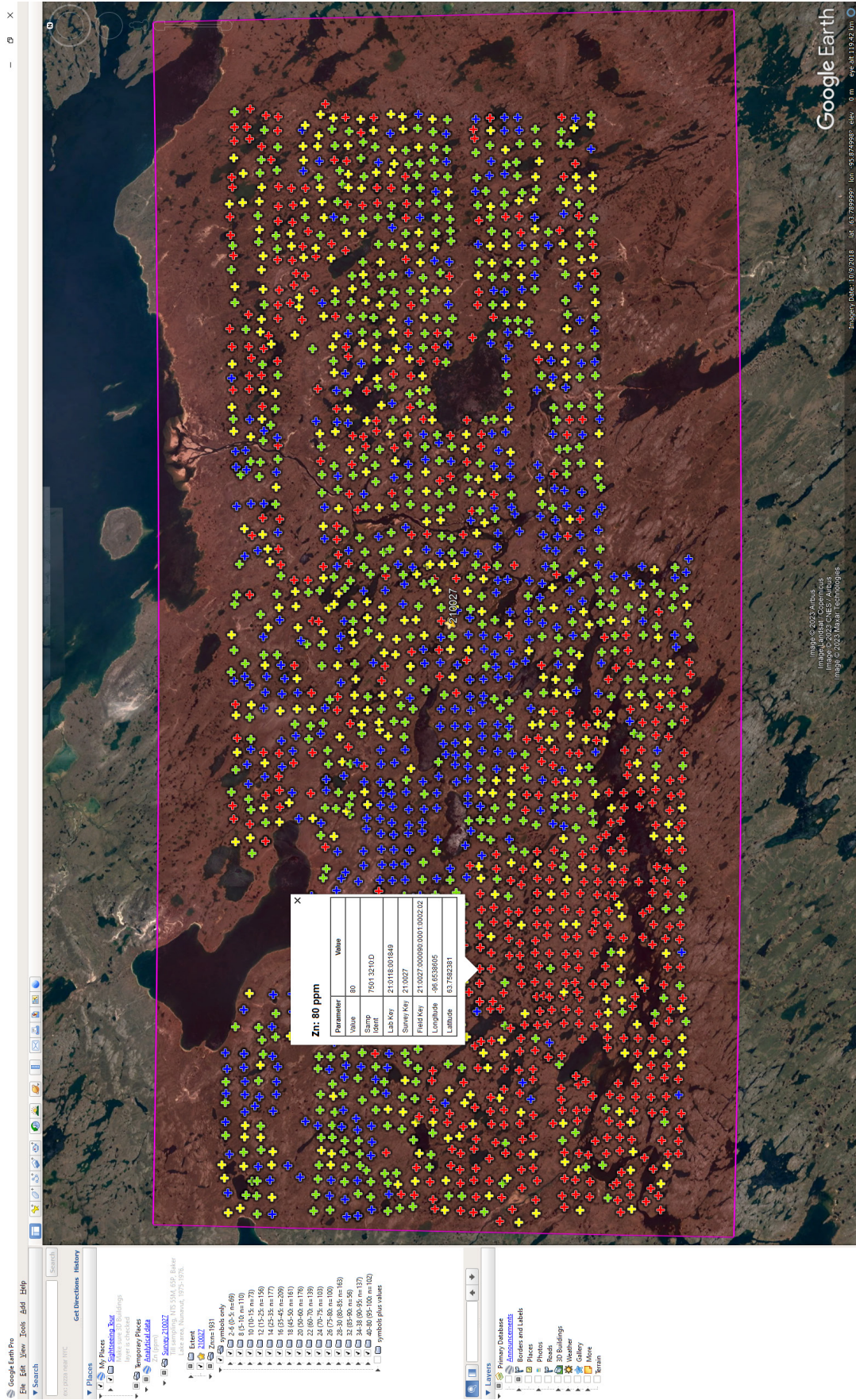


Figure 6. Map of zinc in till samples (<63 µm, ICP-ES, aqua regia digestion), GSC Open File 1335 (https://geochem.nrcan.gc.ca/kml/data/analyses/bdl210118/bdl210118017130100044004_ww_e.kmz)

CHALLENGES

The earlier section placing CDoGS in context identified challenges that are faced by almost every Government-based scientific database activity. Geochemical data face additional challenges in adhering to FAIR principles, centred on the need for better metadata (Chamberlain et al., 2021). A recently published International Union of Geological Sciences (IUGS) manual for establishing a global geochemical database (Demetriades et al., 2022) includes a wealth of valuable advice on how to carry out geochemical surveys, to ensure that the data have long-term value. Two major challenges that are specific to geochemical data are the subject of ongoing development within CDoGS and involve data loading and data integration.

Data loading

Data loading within CDoGS remains a very time-consuming task. There is no standardized workflow for how scientists manage their geochemical data, for sample collection, through preparation and analysis, to publication. Different scientists have different approaches, and use different software. The end result is an enormous variability in the format of the final published data.

A major problem is caused by missing or erroneous metadata. For example, analytical detection limits are frequently not reported, or the values are inconsistent with the actual data. A recent effort spearheaded by M.B. McClenaghan has created a metadata template designed to be included in every GSC-Northern Division Open File that publishes geochemical data (McClenaghan et al., 2020; *see* Campbell et al. (2021) and McCurdy et al. (2023) for examples of its application).

The use of MS Excel[®] as a de-facto standard for publishing geochemical data is also problematic. A single GSC Open File may contain data spread across several different files or different worksheets within a single file. These different worksheets are frequently inconsistent with one another. Sample identifiers may change (e.g. X-YZ-0201 may change to YZ/201). Sample sets may be inconsistent from one worksheet to another (e.g. different numbers of samples). Header rows and column identifiers often need to be edited before data can be transformed into another format. Spreadsheet formatting issues affect the data as received from laboratories and also the data published in Open Files. They could be overcome by a combination of education and well defined guidelines.

Initiatives such as IGSN 2040, which is promoting the idea of a global persistent identifier for geochemical samples (Klump et al., 2021; Lehnert et al., 2021), are critical first steps toward establishing a more robust approach to storing and managing digital geochemical data.

Data integration

In an ideal world, it should be possible to easily combine different geospatial data sets from many different sources into a single integrated data collection. The combined data could then be interpreted using advanced GIS techniques. The different data sets could come from many different sources (geochemistry, geophysics, bedrock geology, topography, hydrology, etc.). At present, combining data sets across disciplines is a very onerous manual task. Standards such as GeoSciML and ISO 19115 will help to reduce the challenge, but their usefulness is critically dependent on the existence of high-quality detailed metadata for the component data sets.

It is still very difficult to integrate different geochemical data sets within CDoGS. The issue of data levelling and integration has been addressed by numerous scientists over the years. Examples include:

- Daneshfar and Cameron (1998) give examples of data levelling using NGR stream sediment data in British Columbia.
- Allard (2004) gives an example using till data from southwest New Brunswick.
- Grunsky (2010) gives a thorough discussion of data levelling challenges.
- Amor et al. (2019) discuss the challenges of integrating lake sediment data across the Labrador Peninsula.

In every example, the levelling was complicated. Automated procedures would appear to be well beyond current capabilities. The need for high-quality metadata are readily apparent, especially with respect to the sample media and the analytical methodologies. Progress is seriously hindered by the absence of any appropriate detailed ontologies.

One beneficial feature of CDoGS is that the database does have a fairly sophisticated methodology for attaching keywords to surveys. The list of keywords is open-ended, and the current set can be viewed at: https://geochem.nrcan.gc.ca/cdops/content/tables/list_kwd_en.htm. The keywords are grouped into nine categories:

- geological,
- analyzed material (hierarchical),
- analyzed quantities,
- analytical techniques (hierarchical),
- decomposition techniques (hierarchical),
- geological/geographic,
- geographic,
- sample preparation procedures, and
- statistical analysis.

Three of these categories have a hierarchical structure. The keywords for analyzed quantities have an additional level of structure, grouping quantities into a handful of categories such as element, oxide, aqueous, etc. Keywords are taken from authoritative sources wherever possible. Geological terms are taken primarily from the GeoRef Thesaurus (Goodman, 2008). Geographic places are taken from the Canadian Geographical Names Database (UNGEGN, 2006). The keyword structure could potentially serve as a starting point for developing a more sophisticated, consensus-based ontology. A useful ontology should be able to address the question of the degree of similarity of two different data sets. One particularly valuable feature would be the identification of analytical methodologies that are essentially identical, perhaps differing only in the lower detection limit, or the precise details of instrumentation (e.g. different pH meters). Cox et al. (2021) give some straightforward guidelines for developing keyword lists which will have long-term value.

CONCLUSION

Developing and maintaining sophisticated data-management systems within government bureaucracies is very difficult. This generalization applies to all levels of government worldwide. Data custodians have a primary responsibility to ensure that the data are not lost, and a secondary responsibility to ensure that they remain accessible. This paper has outlined some of the challenges that CDoGS faces in meeting these responsibilities, and some of the opportunities that may reduce the challenge going forward. The brief overview of CDoGS serves as a demonstration of the value of a comprehensive data-management system.

APPENDIX: BRIEF EXPLANATIONS OF ACRONYMS

Reliable, detailed information on most of the IM/IT acronyms can be readily obtained via Wikipedia.

CanMET: The Mines Branch, Energy, Mines and Resources (EMR, now Natural Resources Canada) was renamed CANMET, the Canada Centre for Mineral and Energy Technology in 1975 (Ignatieff, 1981).

CCRS: Canada Centre for Remote Sensing. Created in February 1971 as a branch of EMR, with Lawrence Morley as the first Director General. It merged in April 1987 with Surveys and Mapping Branch to become the Surveys, Mapping and Remote Sensing Sector (SMRSS). CCRS was merged with Mapping Information Branch (MIB) in July 2013 to become the Canada Centre for Mapping and Earth Observation (CCMEO).

CGDI: Canadian Geospatial Data Infrastructure. Established in 1999 under the GeoConnections program within ESS (GeoConnections, 2012; Hatfield Consultants, 2020).

CGEO: Canadian Group on Earth Observation (*see* IACG).

CGKN: Canadian Geoscience Knowledge Network. A federal-provincial initiative to co-ordinate and improve access to geoscience data across Canada. The program was active from 1999 to 2003 (Broome, 2001; Moore and Buller, 2001; Rupert and Desnoyers, 2001).

CLF: Common look and feel. An approach to website design, which advocates that all of the webpages on a particular site should have the same general appearance and functionality. This helps to 'brand' the website, and also simplifies the end-user's experience, particularly with respect to navigating through the site. TBS is responsible for ensuring that federal government websites comply with its CLF guidelines. A critical component of the current version of CLF is adherence to the WCAG v2.0 standard. Web experience toolkit (WET) is a Treasury Board of Canada Secretariat (TBS) initiative to reduce the burden of compliance.

CSDGM: Content Standard for Digital Geospatial Metadata, developed by the FGDC in the U.S.A. It was adopted by NRCAN in the 1990s, but has since been superseded by the Harmonized North American Profile (HNAP) of ISO 19115.

DCAT: Data Catalog Vocabulary: a resource description framework (RDF) vocabulary designed to facilitate interoperability between data catalogues published on the Web.

DOI: Digital Object Identifier: a persistent identifier used to identify objects uniquely. The DOI for an object (most commonly a document) remains fixed over the object's lifetime, whereas its location and other metadata may change. The DOI system became an ISO standard in 2012.

EMR: Energy, Mines and Resources. A major ministry of the Canadian Government; it superseded the Department of Mines and Technical Surveys in 1966 and was itself superseded in January 1995 by Natural Resources Canada (NRCAN).

ESS: Earth Sciences Sector. A top-level branch of NRCAN, created in August 1995 by the merging of the GSC and Geomatics Canada (the rebranded Surveys, Mapping and Remote Sensing Sector (SMRSS)). It disappeared in a January 2017 reorganization with most of the staff, including all of the GSC, moving into the newly created Lands and Minerals Sector (LMS).

FAIR: Findability, Accessibility, Interoperability and Reusability. A set of guiding principles for scientific data management and stewardship (Wilkinson et al., 2016).

FCGEO: Federal Committee on Geomatics and Earth Observation (*see* IACG).

FGDC: Federal Geographic Data Committee. Operates within the United States federal government. The committee promotes the co-ordinated development, use, sharing, and dissemination of geospatial data on a national basis.

FGP: Federal Geospatial Platform. An initiative to improve access to the geospatial data holdings of the Canadian government, launched by the FCGEO in 2012.

FTP: File Transfer Protocol: a communications protocol for transferring files between computers. Originally developed in the 1970s, it has been superseded by HTTP for many purposes.

GDR: Geoscience Data Repository. Launched in 2003 as a GSC-specific successor to CGKN. The GDR portal was dismantled in 2012, due in large part to difficulties in complying with TBS CLF rules.

GeoSciML: Geoscience Markup Language. A GML Application Schema that can be used to transfer information about geology.

GIS: Geographic Information System. Software dedicated to the manipulation of geospatial data.

GML: Geography Markup Language. An XML notation for encoding geographic data.

GPS: Global Positioning System. A network of satellites that can be used to precisely locate oneself on the Earth's surface.

GSC: Geological Survey of Canada.

HNAP: Harmonized North American Profile. A profile of the ISO 19115 geographic metadata standard, required for compatibility with FGP.

HTML: *see* the entry for XHTML5.

HTML-RDFa: HTML Resource Description Framework in Attributes. A W3C Recommendation that adds a set of attribute-level extensions to HTML for embedding rich metadata within Web documents.

IACG: Inter-Agency Committee on Geomatics. The IACG was created in 1989. Together with the Canadian Group on Earth Observation (CGEO), it was replaced in 2012 with the creation of the Federal Committee on Geomatics and Earth Observation (FCGEO), which brought these two federal coordination committees together (<https://gogeomatics.ca/all-the-geomatics-acronyms-you-wanted-to-know-but-were-too-afraid-to-ask/>).

IGSN: International Generic Sample Number (originally International Geo Sample Number).

IM: Information Management. Used nowadays to refer more-or-less exclusively to digital data, and difficult to separate from IT.

IM/IT: The combined, and difficult-to-separate, fields of IM and IT. Information technology is focused on the physical aspects, whilst IM is focused on the digital data.

ISO: International Standards Organisation.

IT: Information Technology. Refers exclusively to electronic systems, encompassing computers, other electronic devices, and the networks that connect them.

KML: Keyhole Markup Language. An XML notation for storing geographic data. It was developed for use with Google Earth, and became an OGC standard in 2008.

LMS: Lands and Minerals Sector. A top-level branch of NRCAN, created in 2017 by merging most of ESS with a large part of CanMET.

NGR: National Geochemical Reconnaissance.

NRCAN: Natural Resources Canada. A major department of the Canadian government. It was formally created by the passage in 1994 of Bill C-48 "Department of Natural Resources Act." It involved the merging of EMR and Forestry Canada.

OGC: Open Geospatial Consortium (OGC). An international voluntary consensus standards organization, established in 1994.

OpenLayers: An open-source JavaScript library for displaying map data in web browsers.

OWL: Web Ontology Language. A W3C standard for creating ontologies.

QGIS: (previously known as Quantum GIS). A free and open-source desktop GIS application.

R: A programming language and free software environment for statistical computing and graphics.

RDBMS: Relational Database Management System. Software based on the relational model for storing structured data has dominated the database market for many years. There are numerous alternative technologies, both predating and postdating the rise of the relational model.

RDF: Resource Description Framework. A family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model.

RDFa: Resource Description Framework in Attributes. A W3C Recommendation that adds a set of attribute-level extensions to HTML for embedding rich metadata within Web documents.

SMRSS: Surveys, Mapping and Remote Sensing Sector: a top-level branch of EMR, created in April 1987 by merging Surveys and Mapping Branch (SMB) with CCRS. It was “rebranded” as “Geomatics Canada” in June 1994, and subsequently merged with the GSC to become ESS.

SQL: Structured Query Language. A programming language specifically designed for manipulating data within relational databases.

TBS: Treasury Board of Canada Secretariat. The administrative branch of the Treasury Board of Canada. TBS is responsible for requiring all Federal government websites to adhere to its CLF guidelines.

W3C: World Wide Web Consortium. The main international standards organization for the World Wide Web.

WCAG: Web Content Accessibility Guidelines. They are a set of recommendations for making Web content more accessible, primarily for people with disabilities.

WET: Web Experience Toolkit. A sophisticated framework for creating websites that conform to TBS rules.

WMS: Web Map Service. A standard protocol developed by the OGC for serving georeferenced map images over the Internet.

XHTML5: There are several specifications for HTML (hypertext markup language), of which the most recent is HTML5. It is possible to create HTML5 documents that are also valid XML documents. These documents are generally referred to as XHTML5.

XML: Extensible Markup Language. A hierarchical ‘markup language’. A huge number of document types conform to the XML standard, including MS Word[®] and MS Excel[®].

XSD: XML Schema Definition. A schema language that can be used to ensure that XML documents conform to a precisely defined structure.

XSLT: Extensible Stylesheet Language Transformations: A very powerful language for transforming XML files between formats (e.g. transformation of GML to KML).

Z39.50: A communications protocol to allow computers to share metadata. Development began in the 1970s and the protocol continues to be widely used by libraries.

REFERENCES

Adcock, S.W., Spirito, W.A., and Garrett, R.G., 2013. Geochemical data management - issues and solutions; *Geochemistry Exploration Environment Analysis*, v. 13, no. 4, p. 337–348. <https://doi.org/10.1144/geochem2011-084>

Agnew, P., 2017. Geochemistry – state of the art 2017; *in* Proceedings of Exploration 17: Sixth Decennial International Conference on Mineral Exploration, (ed.) V. Tschirhart and M.D. Thomas. <https://geochem.nrcan.gc.ca/cdogs/content/pub/pub03698_e.htm> [accessed July 28, 2023]

Allard, S., 2004. Levelling methods for regional till geochemistry data from southwestern New Brunswick; *in* Geological Investigations in New Brunswick for 2003, (ed.) G.L. Martin; New Brunswick Department of Natural Resources, Minerals, Policy and Planning Division, Mineral Resource Report 2004-4, p.1–20.

Amor, S., McCurdy, M., and Garrett, R., 2019. Creation of an atlas of lake-sediment geochemistry of Western Labrador and Northeastern Quebec; *Geochemistry Exploration Environment Analysis*, v. 19, no. 4, p. 369–393. <https://doi.org/10.1144/geochem2018-061>

Blair, B.B., 2001. GEOSCAN: The GSC Publication Search Engine; *Geolog*, v. 30, Part 1, p. 17.

Blair, B.B., Gillespie, J., and Patey, C., 1993. GEOSCAN: A unique partnership in delivering geoscience information; *Geological Survey of Canada, Forum 1993; Program with Abstracts (Geological Association of Canada)*, p. 7.

Boisvert, E. and Brodaric, B., 2012. GroundWater Markup Language (GWML) – enabling groundwater data interoperability in spatial data infrastructures; *Journal of Hydroinformatics*, v. 14, no. 1, p. 93–107. <https://doi.org/10.2166/hydro.2011.172>

Boisvert, E., Raymond, O., and Sen, M., (2017). OGC Geoscience Markup Language 4.1 (GeoSciML); Open Geospatial Consortium, 247 p. <<https://portal.opengeospatial.org/files/16-008>> (accessed on July 28, 2023)

Brodaric, B., Boisvert, E., Dahlhaus, P., Grellet, S., Kmocho, A., Létourneau, F., Lucido, J., Simons, B., and Wagner, B., 2018. The conceptual schema in geospatial data standard design with application to GroundWaterML2; *Open Geospatial Data, Software and Standards*, v. 3, cit. no. 15. <https://doi.org/10.1186/s40965-018-0058-3>

Broome, J., 2001. The Canadian Geoscience Network (CGKN); *Geolog*, v. 30, Part 1, p. 6.

Campbell, J.E., McMartin, I., McCurdy, M., Godbout, P.-M., Tremblay, T., Normandeau, P.X., and Randour, I., 2021. Field data and till composition in the GEM-2 Rae Glacial Synthesis Activity field areas, Nunavut and Northwest Territories; *Geological Survey of Canada, Open File 8808*, 21 p. <https://doi.org/10.4095/328454>

Chamberlain, K.J., Lehnert, K.A., McIntosh, I.M., Morgan, D.J., and Wörner, G., 2021. Time to change the data culture in geochemistry; *Nature Reviews. Earth & Environment*, v. 2, p. 737–739. <https://doi.org/10.1038/s43017-021-00237-w>

Clarke, A., 2019. Opening the Government of Canada: the Federal Bureaucracy in the digital age; UBC Press, Vancouver, British Columbia; 295 p. <https://doi.org/10.59962/9780774836944>

Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., and Marinescu, M.-C., 2021. Ten simple rules for making a vocabulary FAIR; *PLoS Computational Biology*, v. 17, no. 6, p. e1009041. <https://doi.org/10.1371/journal.pcbi.1009041> [PubMed](https://pubmed.ncbi.nlm.nih.gov/36111111/)

- Daneshfar, B. and Cameron, E.M., 1998. Leveling geochemical data between map sheets; *Journal of Geochemical Exploration*, v. 63, p. 189–201. [https://doi.org/10.1016/S0375-6742\(98\)00015-6](https://doi.org/10.1016/S0375-6742(98)00015-6)
- Demetriades, A., Johnson, C.C., Smith, D.B., Ladenberger, A., Adánez Sanjuan, P., Argyraki, A., Stouraiti, C., de Caritat, P., Knights, K.V., Prieto Rincón, G., and Simubali, G.N. (ed.), (2022). International Union of Geological Sciences manual of standard methods for establishing the global geochemical reference network; IUGS Commission on Global Geochemical Baselines, Athens, Hellenic Republic, Special Publication no. 2, 515 p. <https://doi.org/10.5281/zenodo.7307696>
- Federal Geographic Data Committee (2000). Content standard for digital geospatial metadata workbook (for use with FGDC-STD-001–1998) Version 2.0. <<https://www.fgdc.gov/metadata/csdgm-standard>> [accessed on July 28, 2023]
- Garrett, R.G., Reimann, C., Smith, D.B., and Xie, X., 2008. From geochemical prospecting to international geochemical mapping: a historical overview; *Geochemistry Exploration Environment Analysis*, v. 8, no. 3–4, p. 205–217. <https://doi.org/10.1144/1467-7873/08-174>
- GeoConnections, 2012. Canadian Geospatial Data Infrastructure: vision, mission and roadmap — the way forward; Canadian Geospatial Data Infrastructure, Geological Survey of Canada, Information Product 28e, 20 p. <https://doi.org/10.4095/292417>
- Geological Survey of Canada, 2014. Strategic Plan 2013–2018. <https://publications.gc.ca/collections/collection_2014/rncan-nrcan/M184-3-2014-eng.pdf> [accessed on July 28, 2023]
- Geological Survey of Canada, 2018. Strategic Plan 2018–2023. <https://publications.gc.ca/collections/collection_2019/rncan-nrcan/M184-3-2018-eng.pdf> [accessed on July 28, 2023]
- Goodman, B.A., 2008. *GeoRef Thesaurus*; American Geosciences Institute, Alexandria, Virginia, U.S.A., 818 p.(eleventh edition).
- Grunsky, E.C., 2010. The interpretation of geochemical survey data; *Geochemistry Exploration Environment Analysis*, v. 10, p. 27–74. <https://doi.org/10.1144/1467-7873/09-210>
- Grunsky, E.C. and Broome, J.H., 2001. The Canadian Geoscience Network: a collaborative effort for unified access to data; *Provincial Geologists Journal*, v. 19, p. 105–109.
- Grunsky, E.C., Rupert, J.D., and Williamson, M.A., 2003. Consolidating Canada’s geoscience knowledge program – contributions to the Canadian Geoscience Knowledge Network; *Provincial Geologists Journal*, v. 21, p. 109–111.
- Hatfield Consultants, 2020. Canadian geospatial data infrastructure primer; Canadian Geospatial Data Infrastructure, Geological Survey of Canada, Information Product 60e, 28 pp. <https://doi.org/10.4095/328065>
- Ignatieff, A., 1981. A Canadian research heritage: an historical account of 75 years of federal government research and development in minerals, metals and fuels at the Mines Branch; Canadian Government Publishing Centre, Supply and Services Canada, 353 p.
- Juty, N., Wimalaratne, S.M., Soiland-Reyes, S., Kunze, J., Goble, C.A., and Clark, T., 2020. Unique, persistent, resolvable: identifiers as the foundation of FAIR; *Data Intelligence*, v. 2, p. 30–39. https://doi.org/10.1162/dint_a_00025
- Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, and Wyborn, L. 2021. Towards globally unique identification of physical samples: governance and technical implementation of the IGSN global sample number; *Data Science Journal*, v. 20, no. 1, p. 1–16. <https://doi.org/10.5334/dsj-2021-033>
- Kopf-Johnson, A.G., 1994. GEOSCAN in Proceedings of the 4th International Conference on Geoscience Information (GeoInfo IV), v. 2, (ed.) D.S. Reade, J.C. Caron, A.R. Berger, and A. Barkworth; Geological Survey of Canada, Open File 2315, p. 36–38, <https://doi.org/10.4095/193930>
- Lehnert, K., Klump, J., Ramdeen, S., Wyborn, L., and Haak, L., 2021. IGSN 2040 summary report: defining the future of the IGSN as a global persistent identifier for material samples; Zenodo, 15 p. <https://doi.org/10.5281/zenodo.5118289>
- Loubier, E., 2015. The Federal Geospatial Platform: integrating location into Canada’s public policy through client engagement; Presentation to 2015 INSPIRE / Geospatial world Forum, Portugal. <<https://geospatialworldforum.org/speaker/SpeakersImages/Eric%20Loubier.pdf>> [accessed on July 28, 2023]
- McClenaghan, M.B., Spirito, W.A., Plouffe, A., McMartin, I., Campbell, J.E., Paulen, R.C., Garrett, R.G., Hall, G.E.M., Pelchat, P., and Gauthier, M.S., 2020. Geological Survey of Canada till sampling and analytical protocols: from field to archive, 2019 update; Geological Survey of Canada, Open File 8591, 73 p. <https://doi.org/10.4095/326162>
- McCurdy, M.W., Spirito, W.A., Grunsky, E.C., Day, S.J.A., McNeil, R.J., and Coker, W.B., 2014. The evolution of the Geological Survey of Canada’s regional reconnaissance geochemical drainage sediment and water surveys; *Explore Newsletter for the Association of Applied Geochemists* no. 163, p. 1–10.
- McCurdy, M., Rice, J., Campbell, H.E., and Paulen, R., 2023. Regional lake sediment geochemical data from east-central Labrador (NTS 013-I, 013-J, 013-K 013-N, and 013-O): reanalysis data and QA/QC evaluation; Geological Survey of Canada, Open File 8962, 15 p. <https://doi.org/10.4095/331526>
- Mize, J. 2012. NOAA-NCEI Metadata - ISO 19115:2003 Geographic information - metadata - biological extensions workbook (2.75 MB) - Guide to Implementing ISO 19115:2003(E), the North American Profile (NAP), and ISO 19110 Feature Catalogue with Biological Extensions, NOAA. <<https://repository.oceanbestpractices.org/handle/11329/1281?show=full>> [accessed on July 28, 2023]
- Moellering, H., Brodeur, J., Danko, D.M., Shin, S., and Sussman, R., (2008). The design and intended use of the North American profile V1.2 for spatial metadata. Paper presented at AutoCarto 2008, Shepherdstown, West Virginia, USA. <<https://cartogis.org/docs/proceedings/2008/moellering.pdf>> [accessed on July 28, 2023]
- Moore, A. and Buller, G., 2001. Geochemistry and the Internet: a blueprint; *Geolog*, v. 30, Part 1, p. 15.
- National Information Standards Organization, 2003. Information retrieval (Z39.50): application service definition and protocol specification; ANSI/NISO Z39.50–2003 (maintenance revision of Z39.50–1995), <<https://www.niso.org/publications/ansiniso-z3950-2003-s2014>> [accessed on July 28, 2023]

- Natural Resources Canada, 2015. Guide to harmonize ISO 19115:2003 / North American profile metadata for Government of Canada geospatial data, v. 2.3; June 18, 2015. <https://ftp.geogratis.gc.ca/pub/nrcan_rncan/raster/marine_geoscience/Seismic_Reflection_Scanned/tools/HNAP/HNAP_Conditions%2C_Guidance_and_Examples_v2.3.docx> [accessed on July 28, 2023]
- Paskin, N., 2010. Digital object identifier (DOI®) system; Encyclopedia of Library and Information Sciences; Taylor and Francis Group, London, United Kingdom.; 7 p. (third edition).
- Portele, C., 2016. OpenGIS® geography markup language (GML) encoding standard, version 3.2.2. <<https://repository.oceanbestpractices.org/handle/11329/1120>> [accessed on July 28, 2023]
- Renaud, G., 2004. Metadata and controlled vocabularies in the Government of Canada: a situational analysis. <<https://dcpapers.dublincore.org/pubs/article/view/775/771>> (accessed on 2023-07-28)
- Rupert, J. and Desnoyers, D., 2001. CGKN data discovery tools; Geolog, v. 30, Part 1, p. 7.
- Rupert, J.D., Broome, J.H., and Nolan, L., 2002. The Canadian Geoscience Network: an update on the collaborative effort for unified access to geoscience data; Provincial Geologists Journal, v. 20, p. 111-113.
- Shukle, P., 2014. Presentation to Government Operations Committee, Parliament of Canada, Ottawa, 15th May 2014. <<https://openparliament.ca/committees/government-operations/41-2/26/Prashant-shukle-1/>> [accessed on July 28, 2023]
- Spirito, W.A. and Adcock, S.W., 2010. Canadian geochemical data on the web; Explore Newsletter for the Association of Applied Geochemists no. 149, p. 2-8.
- Spirito, W.A., Adcock, S.W., and Paulen, R.C., 2013. Managing geochemical data: challenges and best practices; in *New Frontiers for Exploration in Glaciated Terrain*, (ed.) R.C. Paulen and M.B. McClenaghan; Geological Survey of Canada, Open File 7374, p. 21-26. <https://doi.org/10.4095/292679>
- Treasury Board of Canada Secretariat, 2013. Guide to the development and maintenance of controlled vocabularies in the Government of Canada, 2nd edition. Treasury Board of Canada Secretariat. <https://geochem.nrcan.gc.ca/ftp/data/publications/pub_10588/controlled_vocabularies_2nd_ed.pdf> [accessed on August 4, 2023]
- United Nations Group of Experts on Geographical Names, 2006. Manual for the national standardization of geographical names; United Nations Group of Experts on Geographical Names, United Nations publication Sales No. E.06.XVII.7, 180 p.
- Warren, H.V., Delavault, R.E., Fletcher, K., and Peterson, G.R., (1971). The copper, zinc and lead content of trout livers as an aid in the search for favourable areas to prospect. In *Geochemical Exploration, Proceedings, 3rd International Geochemical Exploration Symposium, Toronto, April 16-18, 1970*, R.W. Boyle (ed.). The Canadian Institute of Mining and Metallurgy, Special Volume 11, p. 444-450.
- W3C, 2009. W3C semantic web frequently asked questions. <<https://www.w3.org/RDF/FAQ>> [accessed July 28, 2023]
- Wikipedia contributors, 2018. Data catalog vocabulary; Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=Data_Catalog_Vocabulary&oldid=1163161827> [accessed on July 28, 2023]
- Wikipedia contributors, 2019. Schema.org; Wikipedia, The Free Encyclopedia. <<https://en.wikipedia.org/w/index.php?title=Schema.org&oldid=1163700962>> [accessed on July 28, 2023]
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L.O., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., . . . Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship; *Scientific Data*, v. 3, cit. no. 160018. <https://doi.org/10.1038/sdata.2016.18>