DEVELOPING A PROOF-OF-CONCEPT ENVIRONMENTAL GEOCHEMISTRY DATABASE IN CANADA: LESSONS ON METADATA STANDARDS

J. Cole, A. Cleaver, and J. Deriaz
Natural Resources Canada, Ottawa, Canada; Jennifer.Cole@NRCan-RNCan.gc.ca

ABSTRACT

A proof-of-concept database is being developed using publicly available baseline geochemical characterization studies for mining projects in Canada. These studies are submitted during the impact assessment and permitting processes across all jurisdictions and are stored on public online registries as PDF files, constituting a large quantity of "inaccessible" data that is limited in its Reusability. Digital reports from over 70 metal mining projects across Canada have been identified as a source of tabulated static test, kinetic test, and mineralogy data for this proof-of-concept. Data is being extracted, verified, standardized, and aggregated along with comprehensive project and sample metadata such as analytical method, project location, geographical setting, geology, commodity, and proposed mine plan. A common digital format and user interface has been developed to support robust queries that allow end users to compile data across multiple projects targeted by specific search criteria.

To meet new scientific challenges in the era of big data and machine learning, large quantities of data are most beneficial if they are easily accessible, well-documented, and reliable. The FAIR data principles – Findability, Accessibility, Interoperability, Reusability – provide guidance on data management to enhance its utilization beyond the initial purpose.

The proof-of-concept database initiative offers valuable data stewardship learnings for the acid rock drainage and metal leaching community and demonstrates a lack of data standardization and consistent meta(data) documentation. To increase confidence and facilitate integration of these decentralized datasets, the provision of metadata needs to be generous and not limited to specific [end use] needs. Thus, a community-led initiative around data standardization and stewardship is recommended to ensure that future datasets meet FAIR principles and can be easily leveraged for innovative reuse.

Keywords: (meta)data standards, data stewardship, Python, database, environmental geochemistry, Impact Assessment

1.0 INTRODUCTION

A notable number of environmental geochemistry characterization studies are completed to support the evaluation of acid rock drainage (ARD) and metal leaching (ML) potential at new mining projects globally. However, raw data is proprietary and held in private highly fragmented datasets. Further, the available guidance and documentation on standard practice (Price 2009; INAP 2018), has not resulted in consistent data standards (i.e., data format and definition) or minimum information requirements. This results in a valuable data that is currently limited in its reuse by industry, researchers, and stakeholders. Cooperative data stewardship would support the application of advanced data analytics and machine learning to seek innovative solutions to mine waste management (Williams et al. 2015, Vaziri et al. 2021, Salzsauler and Meuzelaar 2021, Meuzelaar et al. 2021).

Large quantities of data are most beneficial if they are easily accessible, well-documented, reliable, systematically structured, and machine readable, with the provision of standardized information (i.e., metadata) essential for repeatability and reuse (Copp et al. 2010; Plana et al. 2019; Chamberlain et al. 2021). The broader global geochemistry research community is working towards a framework that will facilitate the sharing and discovery of geochemical data, by advancing the integration of various data repositories (OneGeochemistry 2022) and hosting ongoing discussions (EGU 2022, Goldschmidt 2022) related to community standards and the future of data stewardship. Miller et al. (2021) and Read et al. (2017) document similar initiatives in the water quality community.

Further, many Earth Science researchers, publishers, and funding bodies are adopting the FAIR principles (Wilkinson et al. 2016): *Findable, Accessible, Interoperable, and Reusable*. These are guiding principles to optimise data management and stewardship and enhance data Reusability (note FAIR components will be highlighted herein by capitalization). FAIR principles facilitate finding and integrating datasets, thus enhancing their value for future use.

For the ARD/ML field, data standards and FAIR principles in data management are being evaluated through a proof-of-concept (POC) initiative for metal mining projects in Canada. Data is sourced from impact assessment (IA) and permitting proposals that are both Findable and Accessible in the public domain. This exercise evaluates the Interoperability and Reusability of data and provides valuable lessons that would support collective data management and stewardship efforts within the community.

2.0 METHDOLOGY: PROOF OF CONCEPT DATABASE DEVELOPMENT

In Canada, designated metal mining projects are subject to IA and permitting processes in accordance with each mining jurisdiction. Geochemical characterization reports included with project proposals are stored on publicly accessible online registries. Data presented in these reports is being targeted for inclusion in a POC database, including project metadata and sample meta(data) for static, tests, kinetic tests, and quantitative mineralogy. Figure 1 summarizes the current database development steps detailed in the following sections. Although quality assurance / quality control (QA/QC) measures are applied throughout the POC workflow, it is only discussed in Section 2.5.

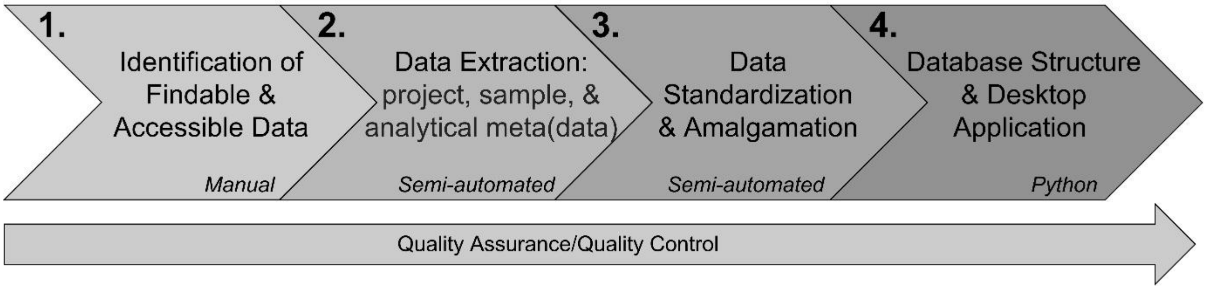

Fig. 1. **Schematic of the POC database development workflow**

2.1 Step 1: Identification of Findable and Accessible Data

2.1.1 Findable and Accessible project reports

Table 1 summarizes the POC database Findable and Accessible project content. Note “proposal” refers to individual IA or permitting submissions whereas “project” is used for an overall mining complex that may include multiple proposals (e.g., expansion, resubmission by new proponent, etc.). The test case comprises 20 precious metal projects and its purpose is to evaluate and refine data standards (Section 2.3) prior to implementing the process for all data-Accessible projects.

Table 1. Findable and Accessible projects within the POC database

Type	Number of Metal Mining Projects	Description
Findable Proposals	154	• IA or permitting proposals that are Findable on public jurisdictional websites
Data-Accessible Projects	72	• Subset of Findable proposals that have Accessible geochemistry reports • May include multiple proposals for same deposit
Test Case Projects	20	• Subset of data-Accessible projects • Used to test and refine data standardization • Contains 8036 unique samples and 17167 individual static test analyses

In total, 154 proposals were Findable on public registries. Many registries do not support Findability, due to regionally organized registries, restricted search capabilities (i.e., advanced filters), unpredictable naming conventions, formal archival procedures, and paper copies available in decentralized locations. Therefore, the list of Findable proposals was compared to NRCan (2022) to address potential discrepancies. The final list targets metal projects that underwent an IA but is not exhaustive of all mines and mining projects in Canada.

Of the Findable projects, 72 had Accessible geochemistry data stored in 130 individual geochemical reports with up to 6 reports per proposal, hindering Findability. Reports date back to 1995 with a peak (36% of proposals) noted between 2010 and 2014 (Figure 2). When reporting historical geochemical testing programs, some projects referenced un-Findable/Accessible reports, impeding report and data Accessibility. Reports were produced by 37 unique author affiliations and 51% can be credited to five consulting firms. Testing was conducted at 22 unique laboratories, with 15% of projects not identifying which one.

Accessible projects span all metal mining regions of Canada (Figure 2). Currently, 10% of the data-Accessible projects are closed or suspended production, 26% are producing mines, 40% are in some stage of permitting and/or development, and the remainder are in exploration or the IA process (Figure 2).

Precious metals comprise 40% of data-Accessible projects, which reflects their leading role in Canadian mineral production value (NRCan 2022), and thus precious metal projects were selected for the test case. Other commodities include cobalt, copper, graphite, iron, lead-zinc, lithium, molybdenum, nickel, platinum group elements, tantalum, and rare earth elements.

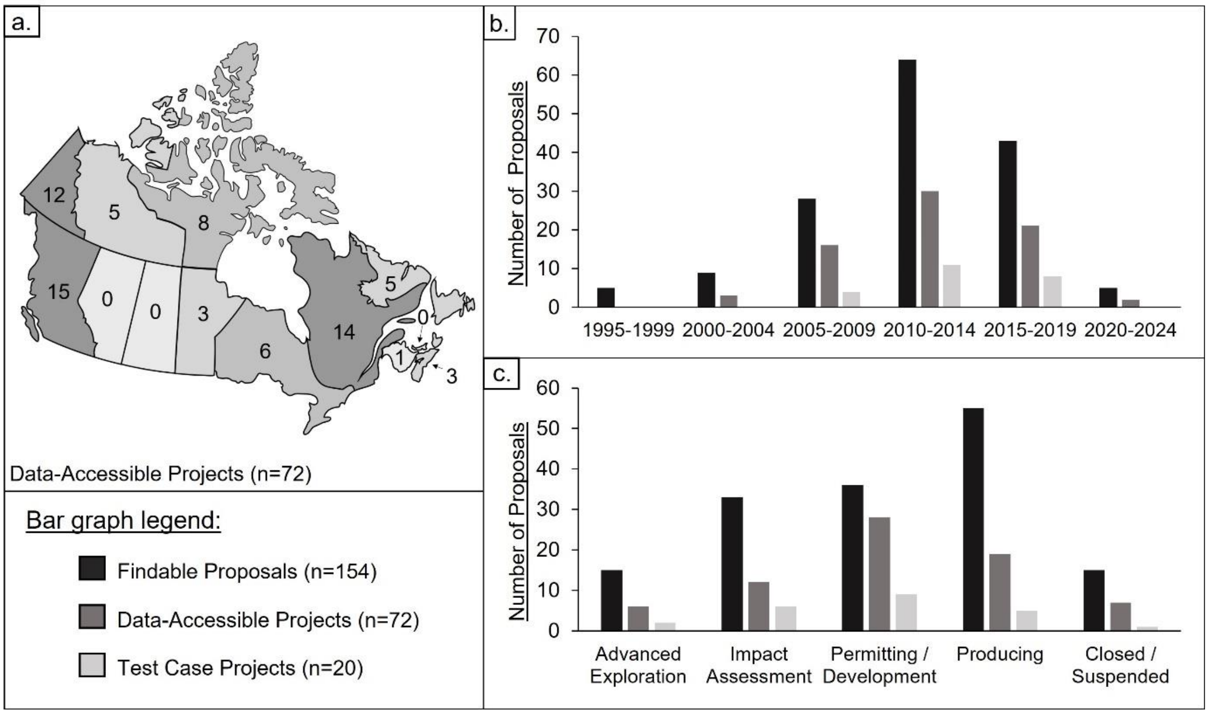

Fig. 2. **a. Project distribution; b. Temporal distribution; c. Current project status**

2.1.2 Findable and Accessible project and sample metadata

Project metadata includes information within the proposal reports, such as the proposal reference and location (e.g., URLs), project location (province, region, and coordinates), commodity, geology, geography (e.g., eco-zone, climate, permafrost type, etc.), and key elements of the proposed operations (e.g., processing method, mine waste management, water treatment, etc.). Some information is not always available within proposals and required supplementary resources, which demonstrates the de-centralized nature of this information, ultimately burdening Findability. For currently producing or closed mines, the project metadata may not reflect the actual operational practices.

Individual sample metadata includes material type (e.g., tailings, mine rock, ore, overburden, etc.), lithology, deposit/zone, borehole, depth, and grain size. However, this metadata was frequently missing, incomplete, or inadequately defined, impeding Accessibility and Reusability. This is discussed further in Section 2.3.2 in terms of metadata standards.

An additional barrier to sample metadata Findability and Accessibility is related to changes in project nomenclature over time (e.g., updates to mine plan, block model interpretation, or changing consultants, etc.) without proper documentation, which limits data Reusability.

2.1.3 Findable and Accessible analytical data

Analytical data includes quantitative mineralogy, whole rock analysis, near-total solid phase elemental analysis referred to henceforth as “trace metals” (e.g., aqua regia, multi-acid digest, etc.), short-term leach tests (e.g., shake flask extraction, synthetic precipitation leaching procedure, etc.), acid base accounting, net acid generation tests, and kinetic tests (laboratory and field). Major impediments to Accessibility include technical report format (e.g., paper copies only, poor quality image-based PDFs, locked PDFs) and data provided in summary statistics tables and/or figures only.

2.2 Step 2: Data Extraction

Project metadata is currently extracted manually through document searches, with future initiatives to consider process automation using artificial intelligence or machine learning.

Tabulated sample metadata and analytical data is extracted from PDF files to Microsoft Excel (MS Excel) using the Camelot library of the Python programming language. Extracted analytical data maintains the original tabular structure in MS Excel and manual verification is conducted for each project to ensure accuracy (Section 2.5).

2.3 Step 3: Data Standardization and Amalgamation

Data standardization enhances the Interoperability of fragmented datasets and facilitates data integration and Reuse in advanced analytics. At a practical level, it supports machine readability, is necessary for data to be Findable within the database, and thus enables successful search queries and data amalgamation across all projects.

Data standards were developed, tested, and refined using the test case projects. The process is semi-automated, with manual verification prior to applying standardized Python dictionaries; future efforts will explore fully automating the process. Data standards are considered for project metadata, sample metadata, analytical methods, and data format.

2.3.1 Project metadata standardization

Project location is documented as reported, including province and region, plus all location data formats are maintained in the POC including geodetic coordinates (latitude and longitude), Universal Transverse Mercator projection (including zone, northing, and easting for the North American Datum), and National Topographic System map). Additional information includes permafrost type, geological province (Natural Resources Canada or NRCan 2019), and terrestrial ecoprovince (Statistics Canada 2021).

Mine operation standards, related to processing methods, mine waste management, and water treatment, were derived from industry references (e.g., INAP 2018; Price 2009) or defined within the project team to obtain consistent nomenclature across projects. Since one purpose of standardized project metadata is to facilitate database queries, the overall objective was to identify simplified and relevant mine plan information only. Details are lost through simplification, especially where multiple methods are proposed, and in some cases, interpretation is required. The original report references are available should detail or data verification be needed by end users.

Standardizing project geology descriptions is also challenging due to complexity of ore deposit geology; Hofstra et al. (2021) is one example of standardized deposit nomenclature that was considered. Applying this approach would require interpretation by geologists unfamiliar with each deposit and constant monitoring for the latest deposit models. Therefore, the database records the most current interpretation of general deposit environment, deposit type, and classification sourced from technical reports and disclosure reporting (i.e., National Instrument 43-101), proponent websites, or other online mining resources. This resulted in approximately

50 unique deposit types within the database and does not currently satisfy Interoperable principles, supporting the need to improve the system.

2.3.2 Sample metadata standardization

Site specific sample metadata (e.g., unique sample identification number (ID), deposit/zone, rock type) is retained in the POC. Samples are defined based on material type including waste rock, ore, quarry/haul road, overburden, and tailings. In addition, the project name is included for each sample to relate the sample data back to project metadata (Section 2.4).

When included, sample location metadata is generally limited to borehole IDs and depth intervals. Only 35% of the 72 data-Accessible projects provide cross sections or block model images to illustrate sample locations within the context of the deposit geology and mine development, while 14% of projects provide example cross sections only to portray "typical" site geology that may include a subset of sample locations. Further, mapping of projected mine rock sample locations at surface is not possible due to a lack of borehole metadata (e.g., coordinate, azimuth, and dip).

Sample lithology nomenclature usage is highly variable and ranges from the use of "rock" to formation names without additional context. In the POC, 449 distinct lithology entries were identified and retained. To facilitate Interoperability, two alternative labels are added using Python dictionaries as follows:

- *high-level grouping of major rock types:* includes 12 organizational groups based on overall chemistry to allow a broader and more simplified search option, including a distinct group for compound lithologies (two or more combined)
- *specific lithology names:* includes 139 unique rock names selected based on the GeoSciML geological data standards (OGC-CGI 2021) with some adaptations for Canadian-accepted nomenclature (e.g., argillite, banded iron formation, etc.).

2.3.3 Analytical method nomenclature standardization

Geochemical test programs vary to meet project objectives, with additional inconsistency in data format and nomenclature introduced by each jurisdiction, practitioner, and laboratory. This results in highly fragmented datasets that can only be Interoperable with the application of data standards. To date, this process has been completed using static test data from the 20-project test-case but will constantly evolve as new projects are added.

Standardized database nomenclature is based on industry references (e.g., INAP 2018; Price 2009) and consensus in the project team. It is assigned after verifying analytical methods by manually reviewing the laboratory certificates and test-specific analytical set up (e.g., type of reactant, volume, and concentration, etc.). This requires addressing inconsistencies in technical jargon and missing (e.g., not Findable) information, resulting in some re-interpretation of available information. Where the method cannot be confirmed, the data was assigned a null or undefined method entry, which significantly limits data Reusability and Interoperability.

While data standardization is complicated by the wide range of accepted nomenclature, the complete omission of or lack of sufficiently detailed sample and analytical metadata renders some geochemical testing programs unrepeatable, thus notably limiting their Interoperability and Reusability. The omissions commonly observed in the POC database are as follows:

- **Trace Metals:** integration of trace metal data must consider the sample digestion methods. Approximately 35% of the test case projects reported incomplete (e.g., “multi-acid”) or no digestion method and are included in an “unknown” method column.

- **Acid-Base Accounting:** these tests provide unique challenges due to a wide range of accepted methods. Significant variability in nomenclature and reporting styles can be standardized using Python dictionaries, however this becomes increasingly difficult when analytical methods are insufficiently described to confidently identify the specific methods used. Consistently observed issues related to acid-base accounting analytes include:

    - **Sulphur species:** Multiple methods to measure sulphate and sulphide produce slightly different results (Price 2009) where approximately 35% of the test case projects did not specify methods or reported unresolvable concentrations with respect to total sulphur.

    - **Carbon species:** incomplete entries such as “CO3” could refer to the analyte (carbonate) or the reported units of concentration (i.e., total carbon as %CO_3). Knowing the specific carbon species and reporting units that were used for carbonate neutralization potential calculations is critical to verify neutralization potential, and calculation checks are often required to confirm the method employed.

    - **Neutralization Potential:** common methods observed include Sobek (1978), modified Sobek (Coastech Research Inc. 1989; Lawrence and Wang 1996), and siderite corrected (Skousen et al. 1997), among others. Various referencing issues were noted, such as “modified Sobek (1978)” or simply “modified Sobek”. In approximately 60% of the test case projects, the methods were not explicitly stated in the report or lab certificates and required some amount of interpretation for incorporation in the database. Subtle method differences (e.g., temperature, test duration, HCl volume and normality, titration endpoint, etc.) were used to identify and/or verify analytical methods, if provided in laboratory certificates of analysis, which frequently repeat the same data omission issues. Otherwise, data was placed in an “NP-unknown” data field.

2.3.4 Data format standardization

Unique data formatting is expected across practitioners and jurisdictions, but the high variability reduces machine-readability and Interoperability. The following were merged into a single consistent formatting style using the Python code:

- **Data format:** number format (e.g., 0.01 vs 0,01; 1000 vs 1 000; scientific format), dates, units, regionally accepted spelling differences (i.e., sulfur vs sulphur), and accent marks.

- **Table format:** transcription errors occurred during data extraction due to merged cells, and sample duplication occurred when data is presented across multiple pages. To ensure each sample is only presented once within a database, results from multiple PDF pages were merged using Python.

- **Detection limits:** values below the reportable detection limit are not always presented in a machine-readable manner. While the Python code can recognize and address “>/<” signs, it cannot detect when it is denoted solely by a change in font format (e.g., text colour, bold, italics, etc.) without the symbol (i.e., 0.1 in place of <0.1).

Python is used to apply data standards and massage extracted analytical data into a consistent structure and format, resulting in one MS Excel file for each project. This approach minimizes human error by reducing the amount of manual data manipulation.

2.4 Step 4: Database Structure and Desktop Application

A schematic of the database structure and desktop application is presented in Figure 3. For Findability within the dataset, the project metadata is keyed to the project name, while analytical data is keyed to the sample ID and associated back to the project name through sample metadata. The desktop application was developed using Python (flask and pandas) to conduct data queries and produce a compiled data file to meet the selected search criteria.

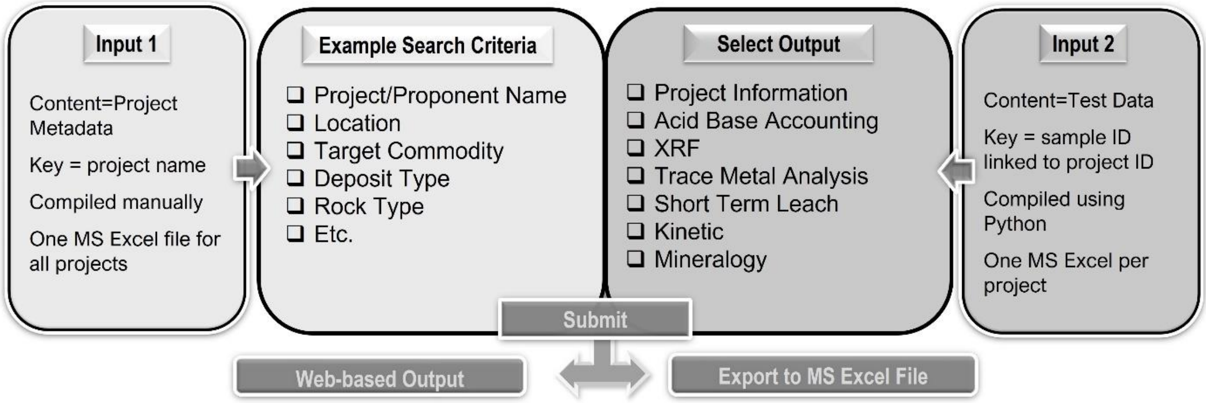

Fig. 3. **Schematic of the POC database structure and desktop application**

2.5 Quality Assurance / Quality Control (QA/QC)

For the POC, analytical data extraction is completed using laboratory certificates of analysis where available, otherwise the tabulated data tables provided within the geochemical characterization reports. Of the 72 data-Accessible projects, only 46% included laboratory certificates of analysis, as this is not a mandatory requirement in Canada. Therefore, for over half of the POC, data accuracy relies directly on the tabulated summary tables provided within the reports. Further, many certificates of analysis are incomplete in terms of analytical metadata to support test repeatability.

This is a critical drawback in terms of data Reusability and confidence by end-users, as many laboratory certificates of analysis provide invaluable information such as internal laboratory QA/QC data (e.g., duplicates, blanks, and standard reference materials, etc.). When lab certificates are absent, the summary tables rarely include analytical QA/QC data and frequently present an incomplete set of analytes that further hinders data validation (e.g., fizz ratings, intermediate pH steps, major ions to support charge balance calculations, etc.). Lastly, summary table data is frequently truncated to a minimal number of significant figures that results in the unintentional reporting of zero values (i.e., 0.00 is reported instead of 0.001).

Therefore, the POC initiative does not evaluate the accuracy or analytical quality of the data presented in either the tables or the laboratory certificates of analysis. This due diligence is left to the end-user to support their specific data Reuse objectives.

QA/QC efforts within the POC focus on ensuring data is extracted and transformed accurately, as provided in the laboratory certificates or report tables. This process was assessed for randomly selected samples representing 10% of the total project dataset for 10 projects. Continual modifications have been made to the Python codes to address shortcomings identified during data verification and, as a result, the code is considered highly reliable.

3.0 DISCUSSION: DATA FAIR-NESS IN ENVIRONEMNTAL GEOCHEMISTRY

Within the global geochemistry community, there is consensus towards adopting FAIR principles and integrating various data repositories (OneGeochemistry 2022). Recent conference sessions (EGU 2022; Goldschmidt 2022) focused on the development of community standards and the future of data stewardship. In their essence, FAIR principles support the Reusability of data and are critical where data is difficult or costly to reproduce, such as environmental geochemistry testing of mine waste materials.

Although the POC database initiative described herein was initiated to support research and IA review, it offers an opportunity to identify how historic and current data management practices measure up to FAIR principles. These learnings are summarized in Table 2 and the observations highlight issues related to analytical and sample metadata standards and common data omissions. Due to the multiple jurisdictional sources of information, proposal and report Findability and Accessibility issues are not addressed in Table 2.

Table 2 also includes preliminary recommendations that could be easily implemented on an individual project basis to increase the Reusability of this valuable data by all end-users. Further, it is recommended that the community consider an initiative to develop consensus on data standards and open access, collective data management practices (e.g., cross-jurisdictional repository) to support discovery and innovation and raise the value of this data for future Reuse.

Table 2. Observations on environmental geochemistry data FAIR-ness

Challenges Observed in POC Database	Recommendations
Findability & Accessibility:	
• PDF and/or data format not machine readable • Data reported in distinct datasets in multiple reports • Sample metadata inconsistent across project reports or sample groups (i.e., historic data) • Insufficient analytical metadata or method description and lack of specific references for repeatability • Data presented as figures or statistics only • Laboratory certificates of analysis not included	• Improved organization and reporting of data in machine readable format • Consistent documentation of project and sample metadata • Provision of laboratory certificates of analysis • Complete description of methods and provision of precise references
Interoperability & Reusability:	
• Variable data reporting formats • Incompatible geology nomenclature • Inconsistent use of analytical nomenclature • Common metadata omissions • Potential limitations to open data and Reuse	• Adoption of geology standard nomenclature • Consensus on minimum information requirements, data format, analytical metadata, and usage license

4.0 CONCLUSIONS

A proof-of-concept database is being developed using a total of 72 Findable projects that include publicly Accessible geochemical characterization reports sourced from IA and permitting proposals in Canada. The amalgamation of mineralogy, static test, and kinetic test data into a single repository has required a significant effort to implement data standards, confirm sample metadata (e.g., location, lithology, etc.), and verify analytical methods.

This demonstrates that FAIR principles (Wilkinson et al. 2016) – *Findable, Accessible, Interoperable, and Reusable* – have not been universally demonstrated in the ARD/ML community thus far. Although substantial data is Accessible, improvements are required for the data to be made Interoperable and Re-usable across these projects.

Collective data management and stewardship within the community, including the application of data standards and minimum information requirements, would provide universal benefit for discovery and innovation, raising the value of the data that has already been collected, and ensuring seamless integration of future data acquisitions.

Further, opportunities are being sought to scale this Canadian initiative to a public-facing database in support of federal open science initiatives (Government of Canada 2022). However, prior to this, consideration must be given to data ownership, license, and Reuse. Moving forward, this type of database is anticipated to be a powerful tool to support mine waste management strategies and mitigate environmental impacts.

5.0 ACKNOWLEDGEMENTS

The authors are grateful to Bill Price, Richard Goulet, Michelle Levesque, Eden Barry, and Sara Ryan for insightful discussions, and to Emmalie Tomini, Jesse Karwaski, Alexandra Korolev, Daniel Cloutier, Rebecca Lord, and Shayna Meinzinger for Python expertise and metadata extraction. This work is supported by CanmetMINING of Natural Resources Canada, the Impact Assessment Agency of Canada, and the Centre for Regulatory Innovation.

6.0 REFERENCES

Chamberlain KJ, Lehnert KA, McIntosh IM, et al. (2021) Time to change the data culture in geochemistry. Nature Reviews Earth & Environment **2**, 737–739.

Coastech Research Inc. (1989) Investigation of Prediction Techniques for Acid Mine Drainage. Mine Environment Neutral Drainage Project 1.16.1.a. November 1989.

Copp JB, Belia E, Hübner C, Thron M, Vanrolleghem P and Rieger L (2010) Towards the automation of water quality monitoring networks. In 'IEEE International Conference on Automation Science and Engineering'. Toronto, Ontario. 21-24 August 2010. pp. 491-496. (IEEE Robotics & Automation Society: New Jersey).

EGU (2022a) Session ESSI3.2 Making Geoanalytical Data FAIR: Managing Data from Field to Laboratory to Archive to Publication. Last accessed 2022-07-11 < https://meetingorganizer.copernicus.org/EGU22/session/42044>

EGU (2022b) Great Debate: Where is my data, where did it come from and how was it obtained? Improving Access to Geoanalytical Research Data. Last Updated 2022-07-06. <https://www.youtube.com/watch?v=nqjpOePQU0w>

Goldschmidt (2022) Session 6gT1: Modern Geochemical Data Systems, Modelling Platforms and Data Standards – creating new opportunities to address scientific problems. Last accessed 2022-07-11. https://conf.goldschmidt.info>

Government of Canada (2022) Open Science. Last updated 2022-02-15. <https://science.gc.ca/eic/site/063.nsf/eng/h_98054.html>

Hofstra A, Lisitsin V, Corriveau L, Paradis S, Peter J, et al. (2021) Deposit classification scheme for the Critical Minerals Mapping Initiative Global Geochemical Database: U.S. Geological Survey Open-File Report 2021–1049, 60 p.

INAP or International Network for Acid Prevention (2018) The Global Acid Rock Drainage Guide Rev 1. Last updated 2018-12-17. <www.gardguide.com>

Lawrence RW and Y Wang (1996) Determination of Neutralization Potential for Acid Rock Drainage Prediction. Mine Environment Neutral Drainage Project 1.16.3. July 1996.

Meuzelaar T, Warren M, Alex A, and Núñez Fernandez P (2021) Breathing new life into old assay data using machine learning methods. Explore: Newsletter for the Association of Applied Geochemists **193**, 1-10.

Miller CB, Cleaver A, Huntsman P, Asemaninejad Al, Rutledge K, Bouwhuis R and Rickwood CJ (2021) Predicting water quality in Canada: Mind the (data) gap. Water Resources Journal of Canada in press.

NRCAN (2019) National Atlas of Canada – Geology and Geosciences: Geological Provinces; Permafrost; and Physiographic Regions. Last updated 2019-06-06 <https://www.nrcan.gc.ca/maps-tools-and-publications/maps/geology-and-geosciences/16876>

NRCan (2022) Minerals and Metals Facts: Minerals and the Economy; Minerals and Mining Interactive Map. Last updated 2022-02-03. <https://www.nrcan.gc.ca/our-natural-resources/minerals-mining/minerals-metals-facts/minerals-and-the-economy/20529#production>

OGC-CGI (2021) GeoSciML Vocabularies Version 4.1. International Union on Geological Sciences – Commission for the Management and Application of Geoscience - Information. Accessed 2021-07-22. <https://cgi-iugs.org/project/geosciml/>

OneGeochemistry (2022) The OneGeochemistry Initiative. Last Accessed 2022-07-14 < https://www.earthchem.org/communities/onegeochemistry/>

Plana Q, Alferes J, Fuks K, Kraft T, Maruejouis T, et al. (2019) Towards a water quality database for raw and validated data with emphasis on structured metadata. Water Quality Research Journal **54**,1-9.

Price WA (2009) Prediction Manual for Drainage Chemistry from Sulphidic Geologic Materials. Mine Environment Neutral Drainage Report 1.20.1. December 2009.

Read EK, Carr L, De Cicco L, Dugan HA, Hanson PC, et al. (2017) Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resources Research **53**, 1735–1745. DOI:10.1002/2016WR019993.

Salzsauler K and Meuzelaar T (2021) Geochemical data, machine learning, and mine planning: using integrated, project-scale datasets for the purpose of environmental and engineering decision making. Abstract in ‘Virtual ICARD 2021 – Meeting Challenge Through Innovation’. 21-23 September 2021. Accessed 21 September 2021. <https://www.icard2021.com/program/>

Skousen J, Renton J, Brown H, Evans P, Leavitt B, Brady K, Cohen L and Ziemkiewicz P (1997) Neutralization Potential of Overburden Samples Containing Siderite. Journal of Environmental Quality **26**, 673-681.

Sobek AA, Schuller WA, Freeman JR, and Smith RM (1978) Field and Laboratory Methods Applicable to Overburdens and Mine Soils. Report EPO-600/2-78-054, U.S. National Technical Information Report PB-280 495.

Statistics Canada (2021) Terrestrial ecozones and ecoprovinces of Canada. Last updated 2021-09-09. <https://www.statcan.gc.ca/en/subjects/standard/environment/elc/2017-map>

Wilkinson MD, Dumontier M, Jan Aalbersberg, I, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 737-739.

Wilkinson MD, Dumontier M, Jan Aalbersberg I, et al. (2019) Addendum: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **6**, 6.

Williams D, Fowler J, and van Zyl D (2015) Mine Planning and Acid Rock Management. In ‘Agreeing on solutions for more sustainable mine water management – Proceedings of the 10th ICARD & IMWA Annual Conference’. Santiago, Chile. 21-24 April 2015. (Eds A Brown, C Buckham, J Burgess, M Carballo, D Castendyk, L Figueroa, L Kirk, V McLemore, J McPhee, M O’Kane, R Seal, J Wiertz, D Williams, W Wilson, C Wolkersdorfer) pp. 1-1935. (GECAMIN: Santiago).

Vaziri V, Sayadi AR, Mousavi A, Parbhakar-Fox A, and Monjezi M (2021) Mathematical modeling for optimized mine waste rock disposal: Establishing more effective acid rock drainage management. Journal of Cleaner Production 288: 125124.