Compiling data from the analysis of surficial materials in Nunavut into a database at the Canada-Nunavut Geoscience Office

T. Tremblay[1] and S. Basso[2]

[1]*Canada-Nunavut Geoscience Office, Iqaluit, Nunavut, tommy.tremblay@canada.ca*

[2]*Canada-Nunavut Geoscience Office, Iqaluit, Nunavut*

Tremblay, T. and Basso, S. 2021: Compiling data from the analysis of surficial materials in Nunavut into a database at the Canada-Nunavut Geoscience Office; *in* Summary of Activities 2020, Canada-Nunavut Geoscience Office, p. 93–100.

Abstract

The Canada-Nunavut Geoscience Office (CNGO) is currently developing a database for data from the analysis of surficial materials in Nunavut. The data preprocessing structure presented in this paper outlines the organization of the geochemical, mineralogical and sedimentological data, which will facilitate its scripted import into a SQLite relational database. This paper focuses on explaining the data preprocessing stage, and the terminology used in the construction of the database.

Introduction

Data from geochemical, mineralogical and sedimentological analyses of surficial materials in Nunavut are available within documents from a variety of sources including the Canada-Nunavut Geoscience Office (CNGO), the Geological Survey of Canada (GSC), academia and the mineral industry. Analyzing geological data from different sources is difficult because the user must make assumptions regarding how comparable the data are; therefore, the complete metadata needs to be catalogued to allow such comparisons between different data. From source to source, and within sources, data are organized in a variety of ways, which makes it difficult to process the data programmatically. Furthermore, in order to succeed in establishing a sound database, the structure and terminology of the data needs to be established; such establishment includes the relationships between sample names, publications, surveys and other published data from the same samples. In particular, naming of the samples is critical, as different names can be used for the same sample in different publications. All this data needs to be correlated. Also, the question of quality assurance–quality control (QA/QC) duplicates and standards must be accurately sorted out, in particular, as to whether the sample is a field or laboratory duplicate. Overall, the metadata attributed to the sample has to be captured as completely and accurately as possible during the preprocessing and processing stages, to allow for complete, successful and precise analyses of the data. The approach should have a solid universal basis, to facilitate its use in the future for data from any type of analysis of any type of geological material.

There is a list, but not a compilation, of numerous surveys and publications (with downloadable datasheets) for Nunavut in the Canadian Database of Geochemical Surveys (Natural Resources Canada, 2020a; see also Adcock et al., 2013; Spirito et al., 2013); this list can be downloaded from GEOSCAN (Natural Resources Canada, 2020b) and the CNGO's website (https://cngo.ca/). However, the Canadian Database of Geochemical Surveys takes a complex approach to capturing data that does not fit well with the CNGO's approach, which requires a more flexible data structure to accommodate a wider variety of data and allow for greater ease in writing scripts. This paper presents the methodology for collecting (herein called preprocessing), aggregating (processing) and analyzing (postprocessing) datasets, including all data and metadata. This paper will focus primarily on the preprocessing stage, which is the most advanced stage to date.

Data terminology and structure

Presented in Figure 1 is the data structure showing the various entities and relationships used to capture almost any type of surficial material analytical data. This structure is the basis for the directory of Microsoft® Excel® files and the SQLite database produced in the preprocessing and processing stages, respectively.

Dataset

The resulting intermediate structure for the surficial material analytical data from a particular document is referred to as a dataset. The name of a dataset is based on the document to which it is related and should be globally unique. For

This publication is also available, free of charge, as colour digital files in Adobe Acrobat® PDF format from the Canada-Nunavut Geoscience Office website: https://cngo.ca/summary-of-activities/2020/.

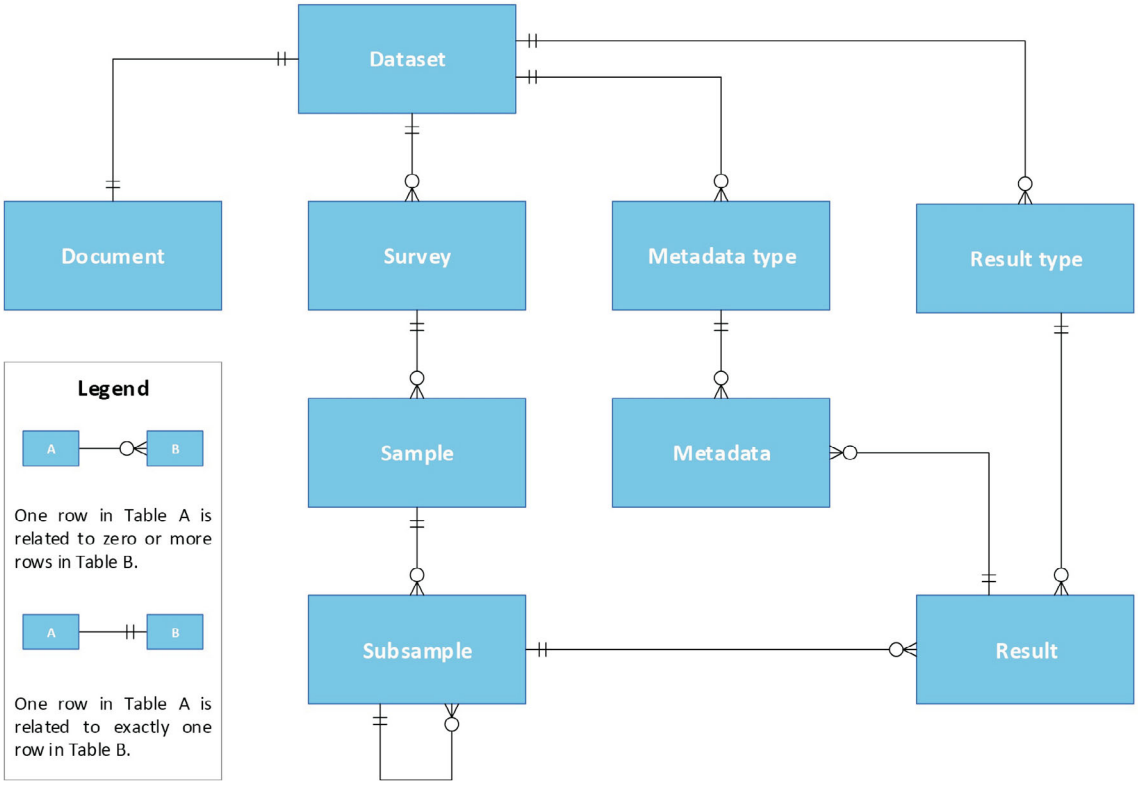

Figure 1: Diagram of data structure for the directory of Microsoft® Excel® files and the SQLite database of analytical data from the surficial materials of Nunavut.

this, reverse domain name notation (Wikipedia, 2020) is used, and this will permit datasets to be uniquely named based on their organization of origin. For example, the dataset name 'ca.gc.nrcan.gsc.OF1575' refers to Open File 1575 from the Geological Survey of Canada of Natural Resources Canada of the Government of Canada.

Document

Documents of interest are those that contain geochemical, mineralogical or sedimentological data; these documents can be published (CNGO datasets, GSC open files, scientific journals, etc.) or unpublished (laboratory analysis data sheets, etc.). A document has a one-to-one relationship with a dataset, meaning a dataset must be related to only one document and vice versa.

Survey

Surveys refer to the specific geological effort within a specific region and at a specific time when samples were collected. A survey is identified by its dataset and title. Multiple datasets may reference the same survey, so care should be taken to ensure that these datasets have the same survey name included. In some cases, a survey might also relate to a specific GSC survey number, which is contained within the Canadian Database of Geochemical Surveys. There can be more than one survey related to a single dataset.

Sample

A sample is geological material collected in the field, and is related to a single survey. A sample is identified by its survey, station name, earth material name (e.g., if there is bedrock present at station 09TIAT001, the sample could have an associated earth material name of 09TIAT001A and if till is also present, it could also have an earth material name of 09TIAT001B) and sample name. A sample also has associated location information (e.g., latitude and longitude) and earth material type.

Subsample

A subsample is named as stated in an analytical report, or as reported in a document. As shown in Figure 1, a subsample has a self-referential relationship allowing for deeper subsampling (e.g., sub-subsample of a subsample), if needed. For example, a sample was collected in the field, subsampled for heavy mineral analysis, and within this subsample, a mineral grain was collected and analyzed for trace geochemistry (sub-subsample).

Metadata type and metadata

Together, the metadata type and metadata entities characterize a specific result. For example, a result could have a metadata type of 'method of analysis' with a value of 'ICP-MS'. The structure allows for results to be associated with values for many different metadata types. The metadata

type entity is simply a lookup table that maps IDs (unique integers) to metadata types serving to normalize the data structure, whereas the metadata entity enables the many-to-many relationship between metadata type and result.

Result type

The result type is the quantity or quality analyzed or observed for a specific subsample (e.g., copper concentration). The result type entity is a lookup table similar to the metadata type entity, but in this case, mapping IDs to result types.

Result

The result is the value for the given result type, subsample or metadata (e.g., 56). There is a one-to-many relationship between the result type entity and the result entity whereas a result can have only one result type.

Methodology

The methodology for data compilation is divided into three stages: preprocessing (preparation of the raw data), processing (creation of a single-file relational database from the preprocessed data) and postprocessing (querying the data). This paper focuses on the preprocessing stage as opposed to the two subsequent stages, which are not as advanced. The overall process will be refined and detailed in future papers.

Preprocessing

Geochemical and other types of geological data typically arrive in multiple formats, with unrelated metadata hidden in reports and tables; these factors necessitate conversion to a common structured format. The preprocessing methodology is aimed at capturing all the necessary information, in a normalized manner, and with an ease of manipulation sufficient to allow a large range of users. The format adopted at the CNGO uses a series of directories and Microsoft Excel files, a convenient format in widespread use. The methodology presented in the following outline is based on the data structure presented earlier in this paper:

1) Creating an appropriately named directory that represents a dataset. As mentioned in the 'Data terminology and structure' section, a dataset is named based on its associated document using reverse domain name notation.
2) Creating a subdirectory named 'ORIGINAL' that contains the original document and any associated data files.
3) Populating the dataset directory with the following required files:
   a) DOCUMENT.xlsx
      This is a mandatory file containing one worksheet with one data row (see Table 1).
   b) SURVEYS.xlsx
      This is a mandatory file containing one worksheet with zero or more data rows (see Table 2). Zero data rows in this file would mean zero data rows in SAMPLES.xlsx and zero data rows in the analysis files.
   c) SAMPLES.xlsx
      This is a mandatory file containing one worksheet with zero or more data rows (see Table 3). Zero data rows in this file would mean zero data rows in the analysis files. Data rows in this file are related to data rows in SURVEYS.xlsx through the use of the survey title as a foreign key. Original geographic co-ordinates must be batch-converted to latitude and longitude with NAD83 as the ellipsoid reference.
4) Populating the dataset directory with zero or more analysis files (e.g., BULK.xlsx, HMC.xlsx or GRAINS.xlsx). These are optional files containing one or more worksheets. These files contain the actual surficial material analytical data. The name of the file refers to the specific subfraction (object) linked to the metadata and results within the table. For example, BULK refers to analysis of the whole material, HMC to the heavy mineral content fraction and GRAINS to individual grains; other subfraction file names could also exist (e.g., if interstitial water is analyzed, this file could be named INTERSTITIAL_WATER.xlsx). Within each file, the composition of the worksheets must adhere to the standard format presented in Table 4.
5) Performing quality control using a Python library in development at the CNGO, which provides a programmed interface to the dataset directory; this interface verifies that all of the aforementioned criteria are met, prior to initiating the next stage (processing).

Processing

Processing involves taking all of the Microsoft Excel files from a single dataset, produced in the previous step, and combining them into an SQLite database, with a single file for each dataset. SQLite has been chosen because of its openness, portability and applicability of popular data analysis programming languages (e.g., Python and R) and GIS software (e.g., ArcGIS and QGIS), and the inherent power of SQLite as a relational database. The advantages are discussed in greater detail in SQLite developers (2020). The adoption of SQLite will allow for better quality control measures due to the inherent benefits of a relational data-

Table 1: An example of a worksheet in DOCUMENT.xlsx.

RECOMMENDED_CITATION
Tremblay, T., Sasseville, C. and Godbout, P.M. 2020: Data table accompanying "Geochemistry and mineralogy of glacial sediments, north of Fury and Hecla Strait, northwestern Baffin Island, Nunavut"; Canada-Nunavut Geoscience Office, Geoscience Data Series GDS2020-001, Microsoft® Excel® file.

Table 2: An example of a worksheet in SURVEYS.xlsx.

TITLE	ORGANIZATION	YEAR_BEGIN	YEAR_END	PARTY_LEADER	DESCRIPTION	GSC_CATALOG_NUMBER
2018, Till sampling in Fury and Hecla area. Canada-Nunavut Geoscience Office.	Canada-Nunavut Geoscience Office	2018	2018	Tommy Tremblay		210292
2018, Till sampling in Isortoq Lake area. Canada-Nunavut Geoscience Office.	Canada-Nunavut Geoscience Office	2018	2018	Tommy Tremblay		210293

Table 3: An example of a worksheet of SAMPLES.xlsx.

SURVEY_TITLE	STATION	EARTHMAT	SAMPLE	LAT_NAD27	LONG_NAD27	LAT_NAD83	LONG_NAD83	X_NAD27	Y_NAD27	X_NAD83	Y_NAD83	ZONE	EARTHMAT_TYPE
2018, Till sampling in Fury and Hecla area. Canada-Nunavut Geoscience Office.	18TIAT115	18TIAT115A	18TIAT115A1			70.334809	-86.178963						diamicton
2018, Till sampling in Isortoq Lake area. Canada-Nunavut Geoscience Office.	18TIAT168	18TIAT168A	18TIAT168A1			70.194582	-76.452538						diamicton

Table 4: An example of a worksheet in HMC.xlsx.

SAMPLE	SUBSAMPLE	METADATA_TYPE	chalcopyrite	pyrite	corundum
		FRACTION_MIN_MM	0.25	0.25	0.25
		FRACTION_MAX_MM	0.5	0.5	0.5
		THRESHOLD			
		UNIT	grains	grains	grains
		METHOD	heavy mineral analysis	heavy mineral analysis	heavy mineral analysis
		LAB_PREPARATION	ODM	ODM	ODM
		YEAR	2018	2018	2018
		PRECONCENTRATION_METHOD	shaking table	shaking table	shaking table
		METHOD_GOLD_GRAINS	shaking table	shaking table	shaking table
		LIQUID_DENSITY	3.3g/cm3	3.3g/cm3	3.3g/cm3
		FERROMAGNETIC_SEPARATION	hand magnet	hand magnet	hand magnet
		PARAMAGNETIC_SEPARATION	yes	yes	yes
		MINERAL_IDENTIFICATION_METHODS	visual picking, SEM	visual picking, SEM	visual picking, SEM
		LAB_PICKING	Overburden Drilling Management	Overburden Drilling Management	Overburden Drilling Management
		NORMALIZATION_METHOD	No	No	No
18TIAT115A1	18TIAT115		0	1	1
18TIAT168A1	18TIAT168		0	0	1

SAMPLE/SUBSAMPLE NAMES
METADATA_TYPE
METADATA
RESULT_TYPE
RESULT

base, which include data integrity (enforcing relational and value constraints), normalization (reduced data redundancy compared to a pile of Excel files), accessibility (querying with SQL) and flexibility (easily restructured to accommodate new requirements). The structure of the database is based on the data structure summarized in Figure 1. Unlike the format in the preprocessing stage, the format in the processing stage will more closely map to the data structure. In the simplest case, one entity in Figure 1 represents one table in the database. In addition to the Python library used to interface with the dataset directory presented in the 'Preprocessing section', another library is in development to interface with the SQLite database. Once completed, these libraries will be used in the development of a script to perform the processing stage.

Postprocessing

The postprocessing stage deals with how to query and report the data, in order to analyze or publish the data. With a SQLite database, this step can be performed using various programming languages and GIS software.

At this point, there are individual SQLite databases, each containing one dataset. A single dataset database can be opened in a user's software or programming language of choice and queried using SQL. With a bit of additional effort, all of the individual SQLite databases can be combined into a single SQLite database, allowing queries across multiple datasets.

With a combined SQLite database, one could execute a query such as:

```
SELECT s.*, ss.*, r.*
  FROM samples s,
       subsamples ss,
       metadata_type mt
       result_type rt,
       metadata m,
       results r
 WHERE s.earthmat_type = "diamicton"
   AND ss.sample_id = s.id
   AND rt.value = "corundum"
   AND m.metadata_type_id = mt.id
   AND mt.value = "FRACTION_MIN_MM"
   AND m.result_id = r.id
   AND m.value = "0.25"
   AND r.subsample_id = ss.id
   AND r.result_type_id = rt.id
   AND r.object_type = "HMC";
```

This query would select all results across all datasets for samples with an earth material type of diamicton, with a result type of corundum, a metadata value of 0.25 for the minimum grain size of the subsample (FRACTION_MIN_MM) and an object type of heavy mineral content (HMC). Once selected, the results can be reported in table format, or as cartographic representations. This data can be published as downloadable files, or as a published or online queryable database file. Postprocessing quality control steps are important, and would include standardizing (e.g., removal of less than symbols affecting numerical values) or normalizing (e.g., dealing with background corrections or missing values) the data.

Using the libraries mentioned in the 'Preprocessing' and 'Processing' sections, a script will be developed that combines the individual SQLite databases (each containing one dataset) into a single SQLite database (one containing all of the datasets) and performs additional quality control checks.

Summary

The Canada-Nunavut Geoscience Office is in the process of developing a database of analytical data for surficial materials in Nunavut. This paper details the preprocessing of geochemical, mineralogical and sedimentological data that will facilitate its scripted import into a relational database (processing stage) and subsequent querying and reporting (postprocessing stage). The latter two stages are currently in the initial stages of development using Python and SQLite. This is an important step in data management for Nunavut, and has numerous important implications for future data collection and program development.

Economic considerations

The Canada-Nunavut Geoscience Office's surficial material analytical data structure is a powerful tool for compiling mineral exploration and environment baseline geological data. The quality of data compilation is key to the value of the data for the intended users, which range from geological researchers to applied geological companies. Once compiled, the surficial materials analytical data can help users to a) find significant geochemical or mineralogical anomalies of specific elements or commodities within a large portion of Nunavut, b) understand the significance of geochemical anomalies found by surficial sediment sampling surveys not included in the database, and c) interpret sedimentary transport processes (e.g., glacial, colluvial, alluvial) as an aid to mineral exploration in surficial sediments.

Acknowledgments

This paper benefited from discussions with A. Plouffe, D. Kerr, I. McMartin, D. Mate, W. Spirito and S. Adcock. Those who assisted with data compilation included R. Baines, T. Rowe, C. Mayer, I. Randour and C. Gilbert. The authors thank S. Adcock for reviewing the manuscript.

Natural Resources Canada, Lands and Minerals Sector contribution 20200676

References

Adcock, S.W., Spirito, W.A. and Garrett, R.G. 2013: Geochemical data management – issues and solutions; Geochemistry: Exploration, Environment, Analysis, v. 13, p. 337–348.

Natural Resources Canada 2020a: Canadian Database of Geochemical Surveys; Natural Resources Canada, URL <https://geochem.nrcan.gc.ca/cdogs/content/main/home_en.htm> [November 2020].

Natural Resources Canada 2020b: GEOSCAN database; Natural Resources Canada, URL <https://geoscan.nrcan.gc.ca> [November 2020].

Spirito, W.A., Adcock, S.W. and Paulen, R. 2013: Managing geochemical data: challenges and best practices; *in* New Frontiers for Exploration in Glaciated Terrain, R.C. Paulen and M.B. McClenaghan (ed.), Geological Survey of Canada, Open File 7374, p. 21–26.

SQLite developers 2020: SQLite; Hipp, Wyrick & Company, Inc., URL <https://www.sqlite.org/appfileformat.html> [November 2020].

Wikipedia 2020: Reverse domain name notation; Wikipedia, URL <https://en.wikipedia.org/wiki/Reverse_domain_name_notation> [November 2020].