Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau

Jessica Norris and Gabe Fortin

BRITISH COLUMBIA
Ministry of Energy, Mines and Petroleum Resources

BRITISH COLUMBIA GEOLOGICAL SURVEY

British Columbia Geological Survey GeoFile 2019-04

Ministry of Energy, Mines and Petroleum Resources
Mines and Mineral Resources Division
British Columbia Geological Survey

Recommendation citation: Norris, J., and Fortin, G., 2019. Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau. British Columbia Ministry of Energy, Mines and Petroleum Resources, British Columbia Geological Survey GeoFile 2019-04, 10p.

Front cover: Subglacial till sample. **Photo by Travis Ferbey.**

Data files for this GeoFile can be downloaded from
https://www2.gov.bc.ca/gov/content/industry/mineral-exploration-mining/british-columbia-geological-survey/publications/geofiles#GF2019-04

Ministry of
Energy, Mines and
Petroleum Resources

Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau

Jessica Norris and Gabe Fortin

Ministry of Energy, Mines and Petroleum Resources
British Columbia Geological Survey
GeoFile 2019-04

Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau

Jessica Norris[1] and Gabe Fortin[1a]

[1] British Columbia Geological Survey, Ministry of Energy, Mines and Petroleum Resources, Victoria, BC, V8W 9N3
[a]corresponding author: Gabe.Fortin@gov.bc.ca

Recommendation citation: Norris, J., and Fortin, G., 2019. Assessment report-sourced surface sediment geochemical database: Development and initial data release from the Interior Plateau. British Columbia Ministry of Energy, Mines and Petroleum Resources, British Columbia Geological Survey GeoFile 2019-04, 10p.

Abstract

ARIS (Assessment Report Index System) is collection of about 36,700 reports that detail results from mineral exploration and development programs. Submitted by the exploration and mining industry to comply with Mineral Tenure Act Regulations and maintain titles in good standing, the reports become an open resource after a one-year confidentiality period. Although containing a wealth of information, the assessment reports are typically submitted in a form (e.g., paper or PDF file) that make working with the data cumbersome. To address this problem, we are developing databases that contain information extracted and digitized from assessment reports, and releasing data in formats that are amenable to evaluation using geospatial and analytical software. This first such database contains surface sediment geochemical data (e.g. soil, silt, stream sediment, and till) from the Interior Plateau, and includes location coordinates, sampling parameters and notes, laboratory information, sample preparation methods, and analytical methods. We use a streamlined process for data compilation, verification, product generation, and distribution. Containing information from about 120 assessment reports, the database includes results from more than 34,000 samples, with 1,446,000 determinations from 13 laboratories. The data released herein (BCGS_GF2019-04.zip) are provided as a GeoPackage, which can be imported directly into most GIS, and as a set of CSV files. Updates to the current surface sediment geochemical database will contain additional geochemical data. Future assessment report-sourced databases will include information from drill holes and geophysical surveys.

Keywords: Geochemical data, assessment reports, soil samples, till samples, digitization, compilation, database, data access, analytical methods, chemical element abundance, Interior Plateau, ARIS

1. Introduction

The British Columbia Geological Survey's (BCGS) Assessment Report Indexing System (ARIS) is a collection of reports documenting mineral exploration and development work in the province, dating back to 1947. Assessment reports provide results of geological, geochemical and geophysical studies, and become an open resource for planning mineral exploration, investment, research, land use, and resource management. To date, ARIS contains about 37,600 reports representing more than $2.8 billion of reported exploration expenditures. Newly available assessment reports are released through the BCGS website monthly.

Although containing a wealth of information, the assessment reports are typically submitted in a form (e.g., paper or PDF file) that render working with the data cumbersome. Specific data, such as sampling parameters and analytical results, must be extracted into formats that are useable for detailed analysis and interpretation. To improve the usability of these data, the BCGS is creating open access databases by extracting and digitizing information from assessment reports. Herein we present the first such database, which includes surface sediment geochemical data (soil, silt, stream sediment, and till) from the Interior Plateau (Fig. 1). The database contains information from more than 120 assessment reports, including results from more than 34,000 samples, with 1,446,000 determinations from 13 laboratories. We provide digital data as a GeoPacakge, which can be imported directly into most GIS (see Appendix 1) and as a set of CSV files (BCGS_GF2019-04.zip).

2. Database structure

The data model, based on that presented by Han et al. (2019), consists of eight tables (Fig. 2; Tables 1-8). The data_sample table is the central table of the database; it contains sample locations and metadata. The data_source_info table links to data_sample through the sample_id and contains details on the source assessment reports. The data_analyte, code_unit, code_method, data_cert, code_prep, and code_lab contain the analyses and all associated metadata for each sample.

Generated from Tables 1-8, additional tables (Tables 9-14)

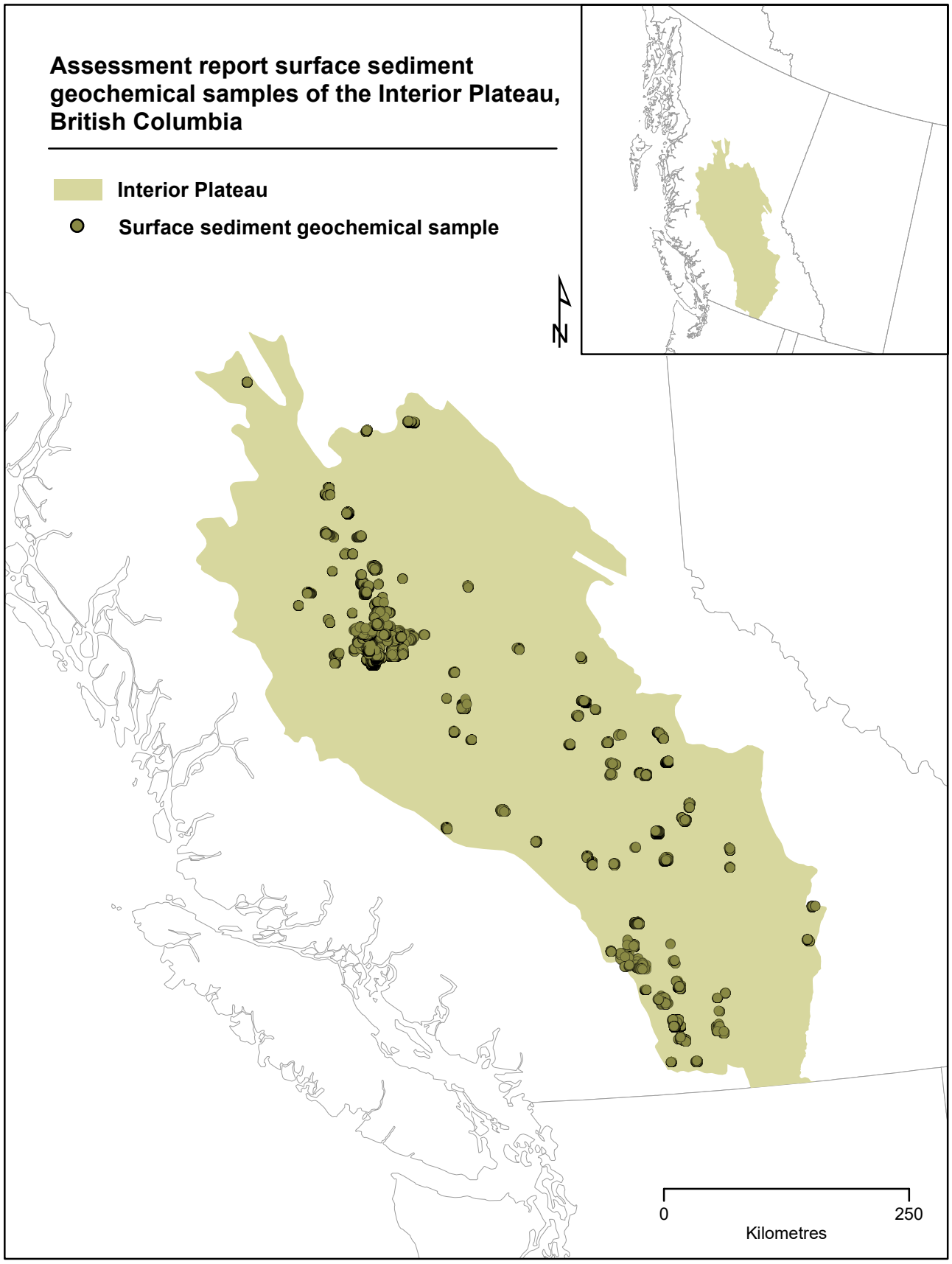

Fig. 1. Assessment report surface sediment geochemical samples from the Interior Plateau.

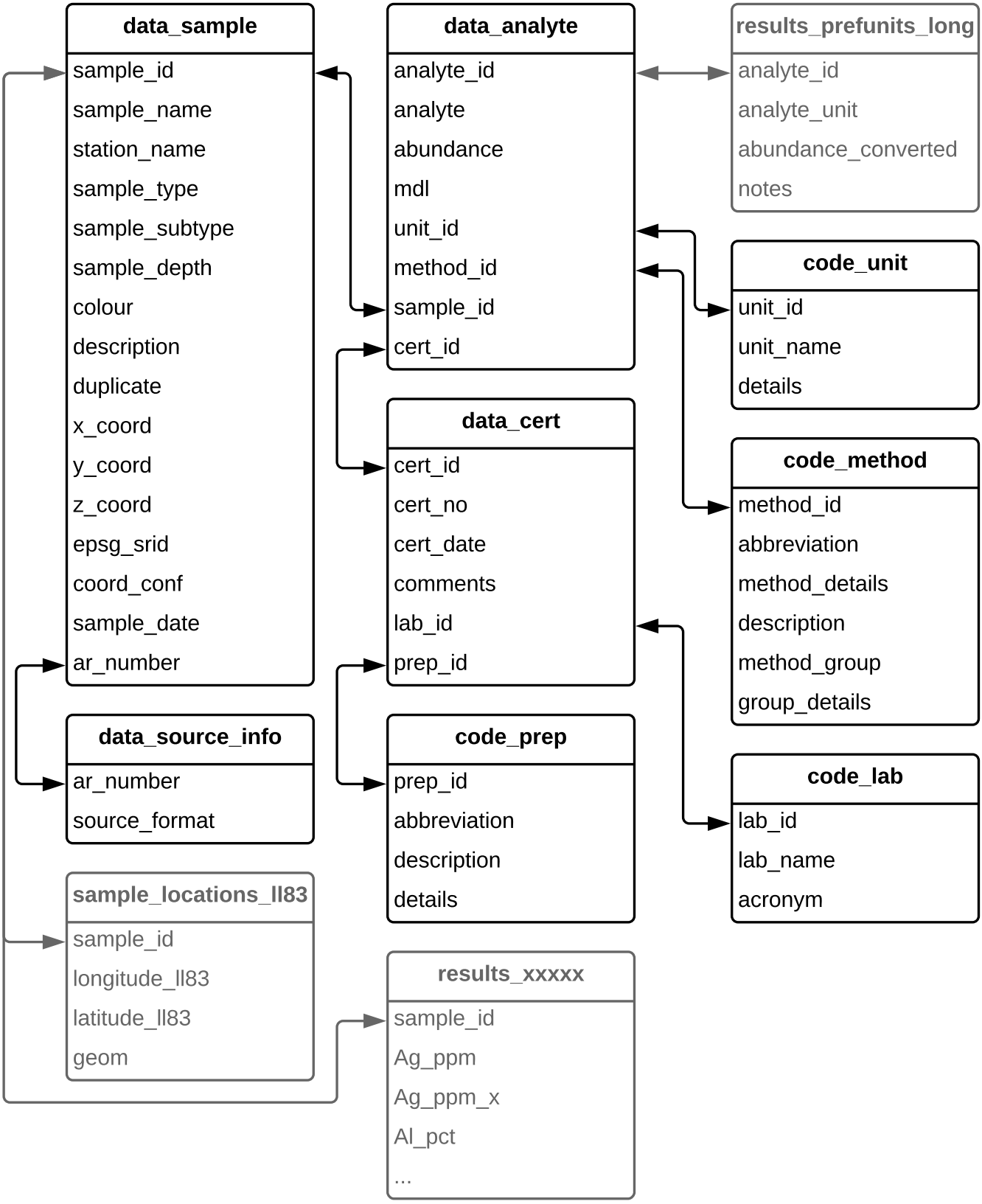

Fig. 2. Database structure: ARIS geochemical data model. Tables in grey are derived products.

Table 1. data_sample: location and metadata for each sample.

Field Name	Type	Description
sample_id	Integer	Unique integer assigned to each sample.
sample_name	Text	Original sample name, as-is from the report.
station_name	Text	When applicable, alternate sample location name from the report
sample_type	Text	Media sampled (e.g. soil, till, etc.), as reported.
sample_subtype	Text	Additional details on sampled media (e.g. A Horizon, B Horizon, etc.).
sample_depth	Text	Sample depth (in cm), as-is from the source report.
colour	Text	Colour of the sampled material, as reported.
description	Text	Description of the sample material or location.
duplicate	Text	Sample name of duplicate sample from the same location
x_coord	Text	X coordinate of sample location, reported in original units as-is from the report
y_coord	Text	Y coordinate reported as-is from the original report
z_coord	Text	Elevation of sample location if reported.
epsg_srid	Integer	European Petroleum Survey Group (EPSG) standard spatial reference system identifier (SRID) of original coordinates.
coord_conf	Text	Confidence in coordinate locations: ‘h’ for high confidence (used for samples where tabular coordinates were in the report); ‘m’ for moderate confidence (used when sample locations had to be digitized from maps; and ‘l’ for low confidence
sample_date	date	Date the sample was collected in the field (YYYY-MM-DD), if reported.
ar_number	Integer	Assessment report number as recorded in ARIS

Table 2. data_analyte: analytical results.

Field name	Type	Description
analyte_id	Integer	Unique integer id for each analytical determination.
analyte	Text	Element symbol
abundance	Text	Reported abundance in original units, prefix ‘-‘ for below detection limit results, ‘>’ for overlimit and ‘NSS’ reported for insufficient sample size. NULL values are excluded from the database as they are taken to indicate no analysis was conducted for the element/sample.
mdl	Text	Minimum detection limit
unit_id	Integer	Links to code_unit table for details on the units in which the results was reported
method_id	Integer	Links to code_method for details on the analytical method
sample_id	Integer	Links to data_sample for sample location and metadata
cert_id	Integer	Links to data_cert for details on the assay certificate the analysis is sourced from.

Table 3. data_source_info: source report information.

Field name	Type	Description
ar_number	Text	Assessment report number as reported in ARIS
source_format	Text	Format of the source report: ‘digital’ for report where digital data (spreadsheets, databases, etc.) were available; ‘georef-ocr’ for reports where sample locations were digitized from georeferenced images and results ocr’d from pdf; and ‘ocr’ for reports where results and locations were extracted from pdf.

Table 4. data_cert: assay certificate metadata.

Field name	Type	Description
cert_id	Integer	Unique integer id for each certificate.
cert_no	Text	Lab-assigned assay certificate number.
cert_date	Date	Certificate date as reported by the lab (YYYY-MM-DD).
comments	Text	Additional notes on the certificate.
lab_id	Integer	Integer id for lab of origin links to code_lab table.
prep_id	Integer	Integer id linking to the prep_method table.

Table 5. code_lab: laboratory metadata.

Field name	Type	Description
lab_id	Integer	Unique integer id for each lab.
lab_name	Text	Name of the lab
acronym	Text	Commonly used acronym for the lab

Table 6. code_method: details on method of analysis.

Field name	Type	Description
method_id	Integer	Unique integer id for each method.
abbreviation	Text	Lab abbreviation for the method/package used
description	Text	Description of the analytical package.
method_details	Text	Additional details on method
method_group	Text	Grouping (coded) assigned to simplify data products
group_details	Text	Grouping of methods

Table 7. code_prep: detail on sample preparation methods.

Field name	Type	Description
prep_id	Integer	Unique integer id for each prep method
abbreviation	Text	Method preparation code from the lab
description	Text	Basic method description
details	Text	Detailed method description

Table 8. code_unit: definition of analyte units.

Field name	Type	Description
unit_id	Integer	Unique integer id for each unit.
unit_name	Text	Name of unit of measurement, common abbreviation or symbol
details	Text	Full name of unit.

are provided in the export (grey in Fig. 2). Results tables follow a naming convention containing two or three parts separated by underscores (Tables 10-12). Table 13 outlines the field naming pattern in an example table. The results_pref_units_long table (Table 14) is created from the data_analyte table; it displays determinations converted to common units for each element.

3. Data compilation

Data compilation follows three steps (Fig. 3, left column): 1) data extraction; 2) data screening (QA/QC); and 3) data loading.

3.1. Data extraction

Data extraction can be broken down into three sub-steps: report selection; OCR (optimized character recognition) or georeferencing; and data staging. For the present release, we only selected assessment reports submitted since 2000 (presumably with samples located accurately), and to those reporting analyses from a certified laboratory.

Because assessment reports are stored as PDF files, we used an OCR (Optical Character Recognition) editor to extract tabular data (e.g., sample names, location coordinates, sampling parameters) and results from analytical certificates to create staging spreadsheet files. All low-confidence characters, as determined from the OCR editor, were compared against the original certificate, and mistakes were corrected manually. In addition, visual checks were made to ensure values were complete and in an appropriate numeric format. Most reports now in the database have tabulated GPS coordinates for sample locations. For a small subset of samples (indicated with an ‘m’ entry in the coord_conf field of the data_sample table), location coordinates were extracted by georeferencing the sample location map and then digitizing each point location.

Table 9. sample_locations_ll83.

Field name	Type	Description
geom	Point	Geometry column that allows points to be plotted directly in GIS.
sample_id	Integer	Sample_id as linked to the data_sample table
longitude_ll83	Decimal	NAD83 (SRID 4269) longitude generated from original coordinates in data_sample, decimal degrees.
latitude_ll83	Decimal	NAD83 (SRID 4269) latitude generated from original coordinates in data_sample, decimal degrees.

Table 10. Result tables naming convention.

results	_fouracid	_halflimits
The results prefix indicates that the table includes determinations.	This portion identifies the analysis method grouping included in the table. (Defined in table 11).	The last portion of the name indicates how determinations below and above detection limit were handled. (Defined in table 12).

Table 11. Analytical method groupings for results_xxxxx_xxxxx tables.

Table Name	Description
ar111	Aqua Regia digestion - 1:1:1 (HNO3:HCl:H2O) with ICP analysis
ar311	Aqua Regia digestion - 3:1 (HC1:HNO3) with ICP analysis
fire	Fire Assay fusion (for Au only) with ICP analysis or AAS or gravimetric finishes
fouracid	Four Acid digestion (HCl, HNO3, HF, HClO4) with ICP or AAS analysis
mmi	Proprietary sample digestion with ICP analysis
unsp	Unspecified/other analytical method

Table 12. Suffixes for results_xxxxx_xxxxx tables.

Table Name	Description
halflimits	Tables with a ‘halflimits’ suffix present results as numbers with determinations below detection limit given as half of the detection limit and determinations over the upper detection limit given as the upper detection limit.
limnull	Tables with a ‘limnull’ suffix present all results as numbers with under and over limit determinations returned as null values.
(blank/absent)	Tables without a second underscore present results in text format values as-is from the original certificate (i.e. including < and > for under and over limit determinations).

Table 13. results_xxxxx example and field naming pattern definition.

Field name	Type	Description
sample_id	Text	Unique sample_id, links to data_sample table.
Ag_ppm	Decimal/Text	Analytical determination for Ag in ppm. Value is reported either as text of decimal as per the table name suffix described above.
Ag_ppm_x	Decimal/Text	Analytical determination for Ag in ppm. Value is reported either as text of decimal as per the table name suffix described above. The ‘_x’ indicates there were multiple determinations for Ag for this sample.
…	Decimal/Text	Analytical determinations for all elements are reported using the above pattern.

Table 14. results_pref_units_long.

Field name	Type	Description
analyte_id	Integer	Links to analyte_id in the data_analyte table.
analyte_unit	Text	Concatenation of the element symbol and units (e.g. Ag_ppm)
abundance_converted	Text	Results from the abundance field are converted to the preferred unit. Reported as text to allow < and > symbols for under and over limit determinations.
notes	Text	Currently only used to denote below detection limit and over limit determinations.

3.2. Data screening

Python scripts are used to check contents of the staging files for accuracy. As used herein, accuracy refers to ensuring the extracted data match what is in the report; it does not imply validation of sampling or analytical methods. More than 30 different verification checks (QA/QC; tables 15 to 17) are run during screening. If any discrepancies are identified, the spreadsheets are inspected and compared to the PDF original to correct them. In cases where an error was determined to have come from the original report (rather than the OCR process), changes were made only where both the error and the correction were clear in the report (e.g., reversal of easting and northing).

3.3. Data loading

Once data have passed the screening phase, they are loaded into a spatial relational database (PostgreSQL/PostGIS) using Python scripts.

4. Products

Data products are generated from the database using a combination of Python scripts and database views. The use of a spatial relational database to store the data makes it simple to generate multiple products. Using record attributes or locations, samples can be linked to other BCGS databases.

Data from the ARIS surface sediment geochemical database

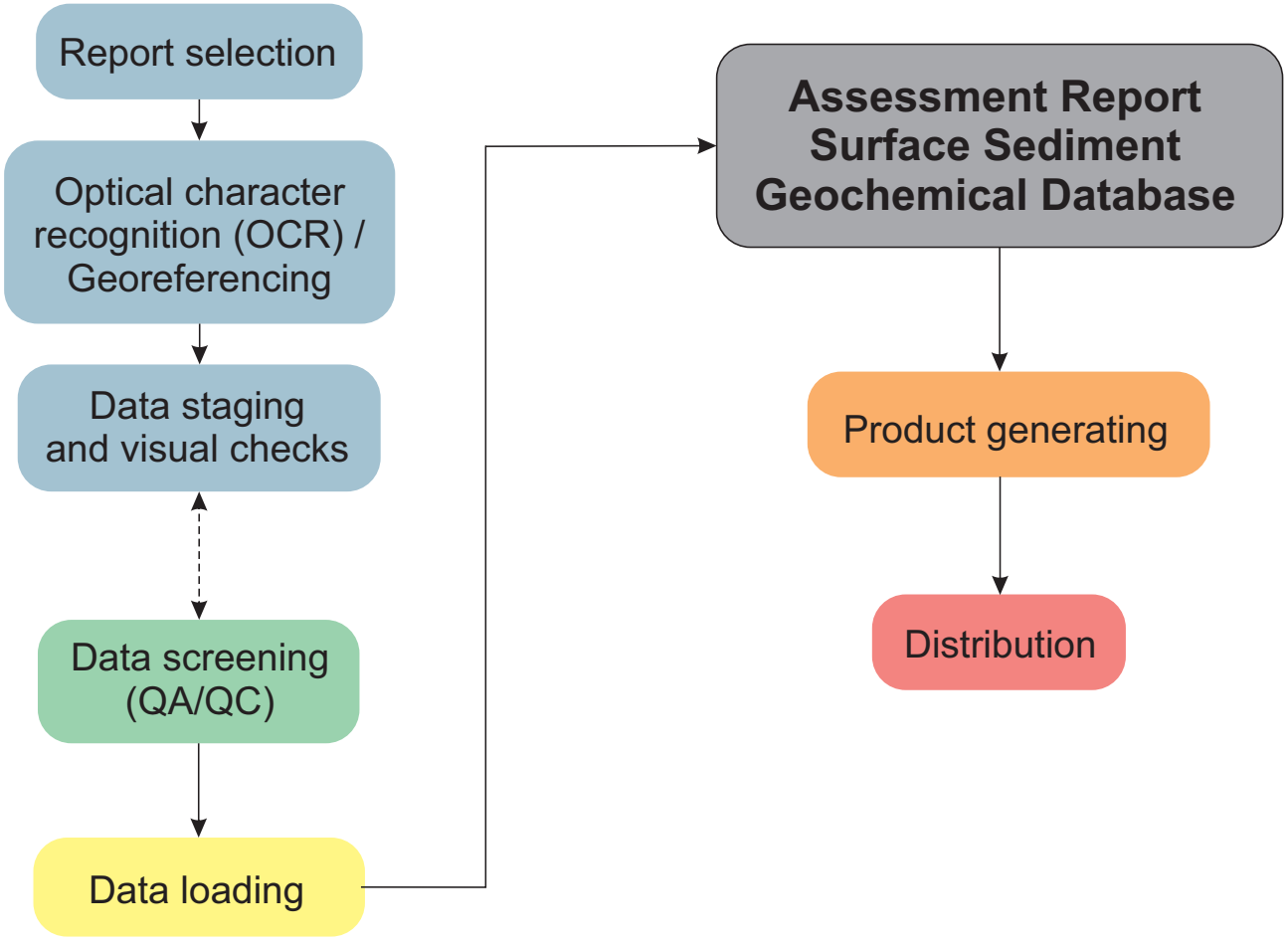

Fig. 3. Data flow diagram.

Table 15. data_sample data checks.

Field(s)	Check
x_coord, y_coord, z_coord, epsg_srid	All entries must be numbers.
z_coord	Must be greater than 0, flagged for manual verification if greater than 3000.
epsg_srid	Must be one of the accepted SRID's known to be valid for BC.
sample_type	Must be one of the types in code_sample_types.
sample_subtype	Must be one of the subtypes in code_sample_subtypes.
coord_conf	Must be one of: l, m or h
x_coord, y_coord, epsg_srid	Combined these must describe a location within BC.
sample_date	Must be a properly formatted, specific date (YYYY-MM-DD).
x_coord, y_coord, epsg_srid	Sample locations that are not within 10km of the report location in ARIS are flagged to be checked manually.
sample_name	Checked for duplicate sample names.

Table 16. data_cert checks.

Field(s)	Check
lab_id	Must match a lab_id in the code_lab table.
prep_id	Must match a prep id in the code_prep table.
cert_date	Must be a properly formatted date (YYYY-MM-DD).
cert_date	Date must be present for all certificates.
cert_no	Certificate number must be present.
cert_no	Certificate number must be unique (data_cert) only allows one entry per certificate.

Table 17. data_analyte checks.

Field(s)	Check
analyte	Only proper element abbreviations are accepted.
unit_id	Units must match a unit from the code_unit table.
mdl	Flagged for manual check if detection limit is missing.
mdl	Must be a properly formatted number.
method_id	Must match a method in the code_method table.
analyte and method_id	Flag entries with 2 results for the same analyte and the same method for manual verification.
abundance	0 values not allowed.
abundance	Must be numeric unless it is an overlimit or below detection limit result.

are made available in two formats: 1) a GeoPackage database including all of the data tables as well as additional tables created from the data; and 2) A set of CSV files (same tables as the GeoPackage file). Additional information on the GeoPackage format and basic use examples are provided in Appendix 1. In the GeoPackage, only the sample_locations_ll83 table is a spatial table. It can be brought directly into most GIS applications; sample information and results can then be joined to this layer as needed. Data can also be viewed on MapPlace 2, the BCGS geospatial web service.

5). Conclusion

The surface sediment geochemical database makes information locked in PDF or paper copy assessment reports, such as from the Interior Plateau, more accessible. Data extracted and digitized from these reports are released as a GeoPackage (BCGS_GF2019-04.zip), which can be imported directly into most GIS, and a set of CSV files. The database continues to grow as more data from the Interior Plateau region are added and as the areal coverage expands. Additional assessment report-sourced databases will include information from drill holes and geophysical surveys. The BCGS invites submission of data in raw and tabular format to ARIS.digital@gov.bc.ca. We would appreciate being notified of errors in the database.

Acknowledgments

We are particularly grateful to those in the mineral exploration industry who supplied data in .xlsx format, and we thank Tian Han for discussions.

References cited

Han, T., Rukhlov, A.S., Riddell, J.M., and Ferbey, T., 2019. A skeleton data model for geochemical databases at the British Columbia Geological Survey. In: Geological Fieldwork 2018, British Columbia Ministry of Energy, Mines and Petroleum Resources, British Columbia Geological Survey Paper 2019-01, pp. 125-135.

Appendix 1. Using GeoPackage files

GeoPackage was chosen as the main delivery format for the ARIS surface sediment geochemical database because it provides a good balance between simple GIS use and more advanced database functionality. It is similar to ESRI's file geodatabase format but is open source, standards based and platform independent. Furthermore, it overcomes many of the limitations of the Shapefile (see http://switchfromshapefile.org) format but maintains its simplicity.

A.1. Opening in QGIS

QGIS is a capable, free, open-source and cross-platform GIS application that can be used to create, visualize and manipulate spatial data from most file or database sources (see https://qgis.org). QGIS natively supports the GeoPackage format. To add layers or tables from a GeoPackage file to a project open the 'Add Vector Layer…' dialog in the 'Layer' menu (Fig. A-1-1) then 'Add Layer'.

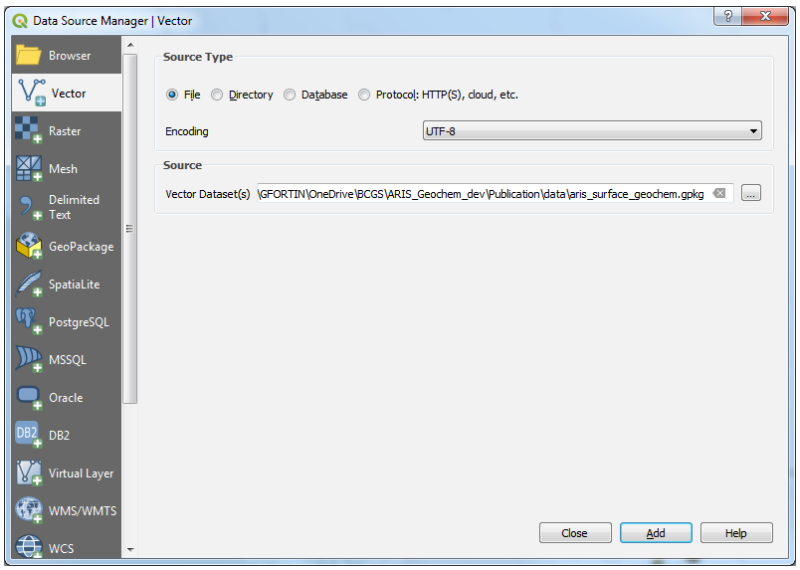

Fig. A-1-1. QGIS Data Source Manager.

Leave 'Source Type' as 'File' and use the '…' button to navigate to and select the GeoPackage file and click 'Add'. In the 'Select vector Layers to Add' dialog (Fig. A-1-2), select all of the layers and tables to be added to the project. Multiple layers can be selected using Shift-Click or Ctrl-Click. Once layers are selected, clicking 'OK' will add them to the project and close the dialog. The 'Data Source Manager' dialog can now be closed.

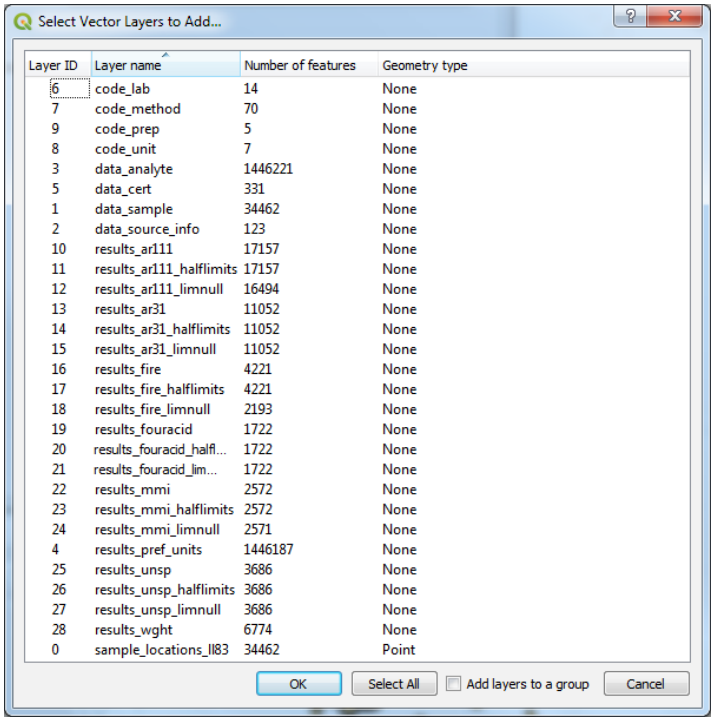

Fig. A-1-2. Select Vector Layers to Add dialog.

A.1.2. Opening in ArcGIS

To add layers or tables from a GeoPackage file to a project in ArcMap, open the 'Add Data' dialog (Fig. A-1-3). Navigate to the .gpkg file and select the layers and tables (Shift- or Ctrl-Click to select multiple) then click 'Add'.

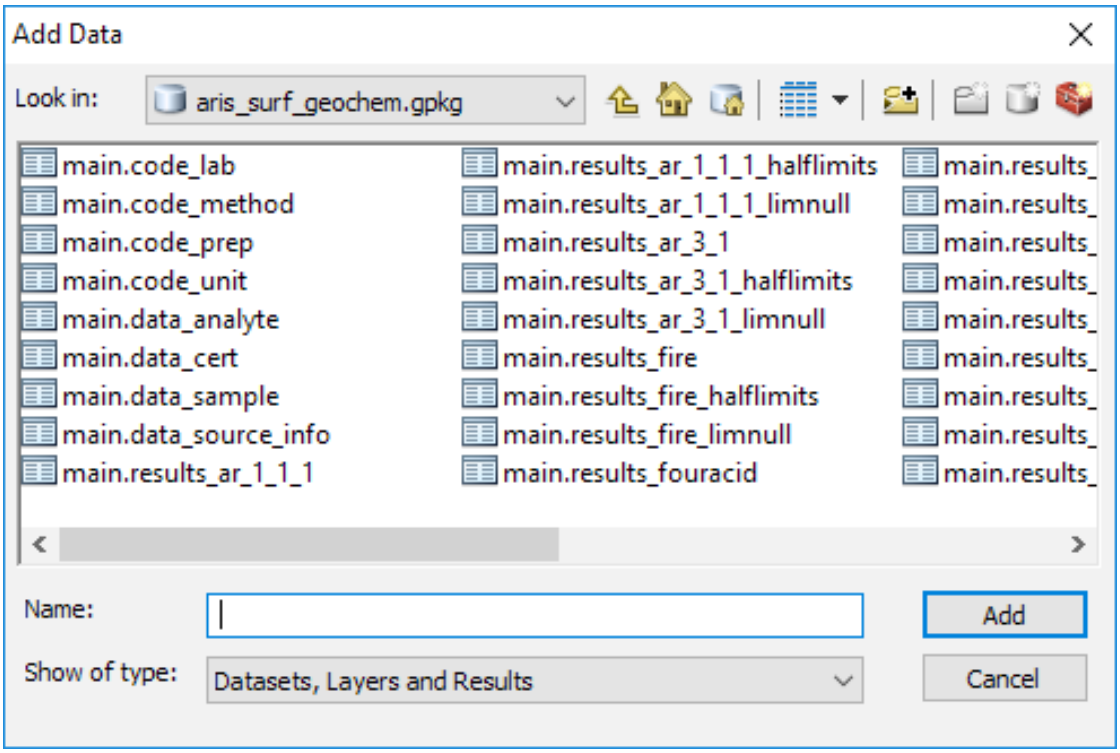

Fig. A-1-3. ArcGIS Add Data dialog.

A.1.3. Using free DB Browser for SQLite

DB Browser for SQLite is a free open source viewer that can be used to explore, manipulate and export data from a GeoPackage. It can be downloaded from https://sqlitebrowser.org. To open a GeoPackage, click the 'Open Database' button, in the 'Choose a database file' prompt, change the entry above the 'Open' button to 'All files (*)' and navigate to the folder containing the GeoPackage (Fig. A-1-4).

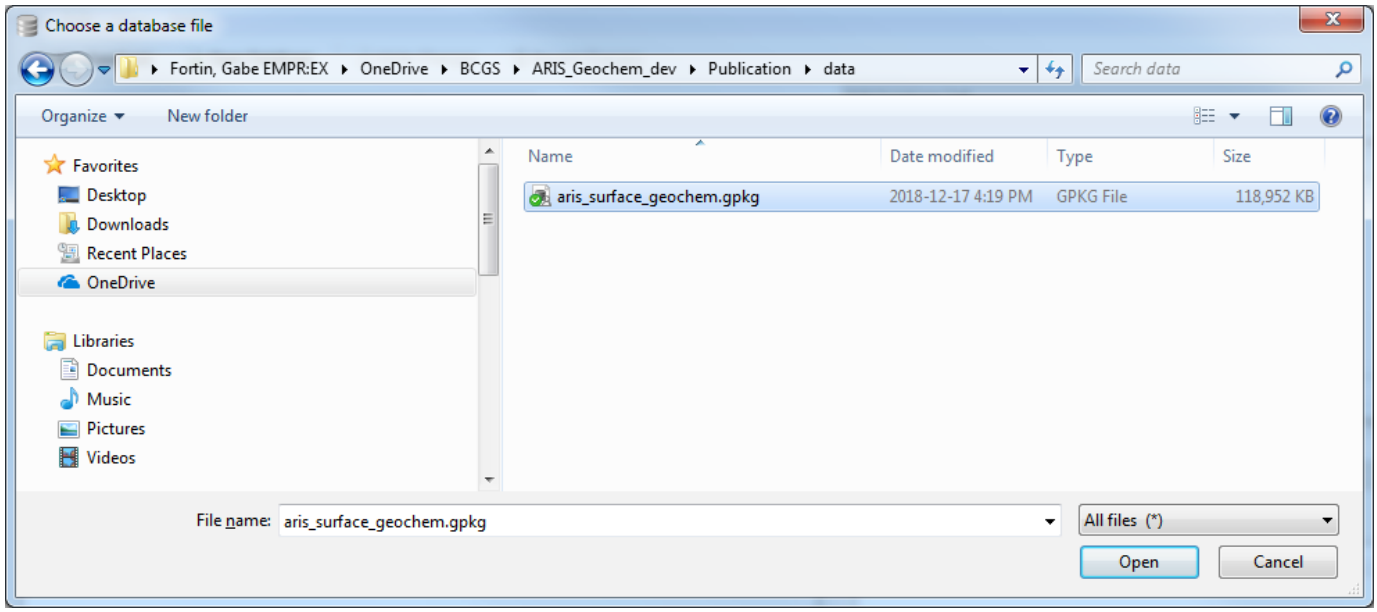

Fig. A-1-4. DB Browser file selection dialog.

British Columbia Geological Survey
Ministry of Energy, Mines and Petroleum Resources