New British Columbia lithogeochemical database: Development and preliminary data release

Tian Han, Alexei S. Rukhlov, Mustafa Naziri, and Adrienne Moy

BRITISH COLUMBIA | Ministry of Energy and Mines

BRITISH COLUMBIA GEOLOGICAL SURVEY

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

New British Columbia lithogeochemical database: Development and preliminary data release

Tian Han
Alexei S. Rukhlov
Mustafa Naziri
Adrienne Moy

New British Columbia lithogeochemical database: Development and preliminary data release

Tian Han[1a], Alexei S. Rukhlov[1], Mustafa Naziri[1], and Adrienne Moy[1]

[1]British Columbia Geological Survey, Ministry of Energy and Mines, Victoria, B.C., V8W 9N3
[a]corresponding author: tian.han@gov.bc.ca

Abstract

A new database system was developed to maintain, update, and manage province-wide lithogeochemical data collected by British Columbia Geological Survey (BCGS) geoscientists and their research partners. The database is populated with data compiled from 214 BCGS publications released between 1986 and 2015. It includes ~11,000 samples with about 250,000 determinations analyzed by 26 analytical methods in 21 laboratories. The database operates through a semi-automated process for data compilation, quality control, upload, and product generation, which helps provide consistent, standardized, and long-term data storage and update, and regular release. The data product derived from the database provided in this release is in a user-friendly format (CVS), which can be easily brought into GIS and geochemical data evaluation software for visualization and analysis.

Keywords: Lithogeochemical database, logical data model, quality control, data product, mineral exploration

1. Introduction

The British Columbia Geological Survey (BCGS) is the custodian for several provincial geological datasets, including bedrock geology, surficial geology, and geochemistry. These datasets and the products derived from them provide geological information for mineral exploration, environmental assessment, and land use planning. However, the lithogeochemical dataset has not been used to its full potential, due primarily to its intrinsic complexity (diverse analytical methods and numerous analytes), scattered data storage, lack of consistent data updates, and inadequate data quality control (QA/QC).

To address these problems, we developed a new lithogeochemical database and initiated systematic compilation of publicly available provincial lithogeochemical data, following release of the new provincial Regional Geochemical Survey (RGS) database (Rukhlov and Naziri, 2015). We: 1) developed a logical data model based on the intrinsic relationships between basic data entities and essential attributes of the lithogeochemical data; 2) systematically populated and updated the database with data compiled from BCGS publications released between 1986 and 2015; 3) conducted quality assurance and control to the existing and newly added data (referring to the original lab reports where available); 4) enabled production of customizable data products derived from the database; and 5) automated in-bound and out-bound data flow for loading the database and generating products. With the above renovations, we established a streamlined process and

an operational database system to manage, maintain, update, and distribute province-wide lithogeochemical data routinely and consistently. This GeoFile is a preliminary release of the lithogeochemical data stored in the new British Columbia lithogeochemical database.

2. Highlights

The lithogeochemical data released herein includes data from 214 papers and reports published by BCGS geoscientists and research partners from the Geological Survey of Canada (GSC) and universities between 1986 and 2015 (Fig. 1). The dataset consists of ~11,000 samples with about 250,000 determinations analyzed by 26 different methods (Fig. 2) in 21 laboratories. The data released are GIS ready, and can be conveniently plotted as an ESRI point shapefile or File Geodatabase (where each point represents a sample location) with the provided datum and projection. The point shapefile was used to prepare Fig. 3, which shows the spatial distribution and coverage of sample locations.

Because sample locations were reported with different spatial accuracies, we introduced an attribute (Coord_Confidence) to account for uncertainties in location (Fig. 3; L = low accuracy, M = moderate accuracy; L = high accuracy). Data users should exercise caution working with the samples with a low accuracy designation. During data QA/QC, such samples were found with the same coordinates as other samples or lacking significant figures.
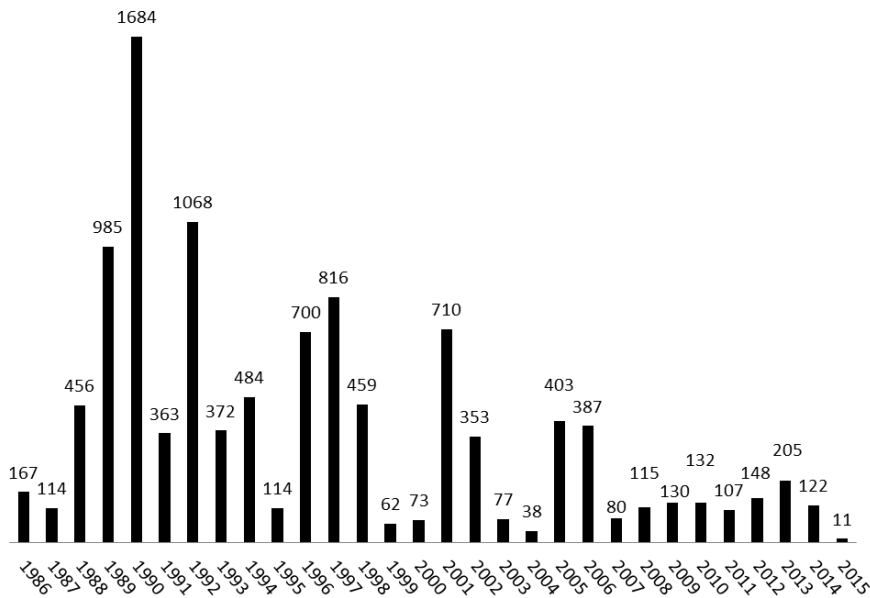
Fig. 1. Number of samples analyzed between 1986 and 2015.
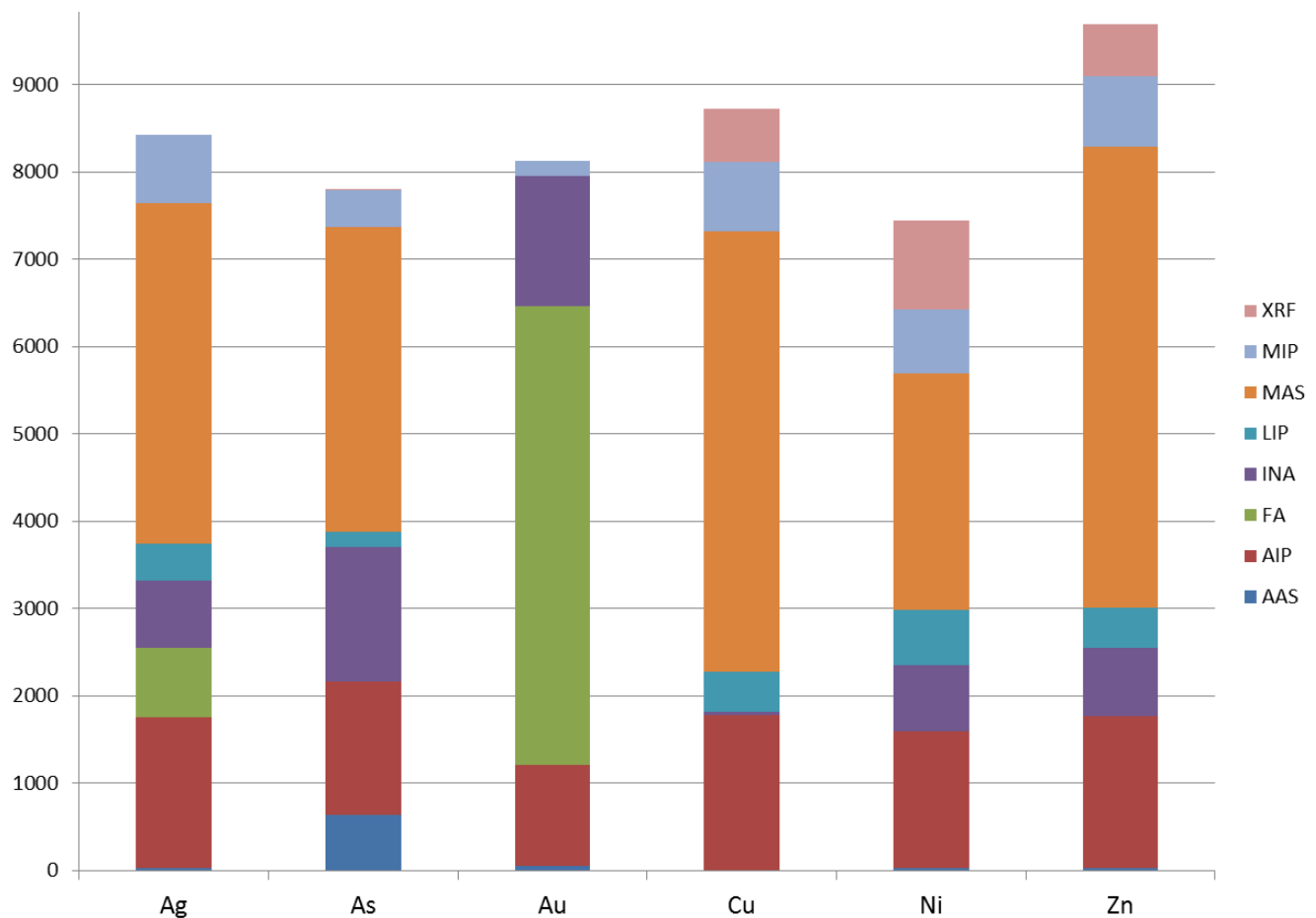
Fig. 2. Major analytical methods for some selected analytes (where XRF stands for x-ray fluorescence; MIP for mixed acid digestion with inductively coupled plasma emission or mass spectrometry; MAS for mixed acid digestion with atomic absorption spectrometry; LIP for lithium metaborate fusion with inductively coupled plasma emission or mass spectrometry; INA for instrumental neutron activation; FA for fire assay; AIP for aqua regia digestion with inductively coupled plasma emission or mass spectrometry; and AAS for aqua regia atomic absorption spectrometry).
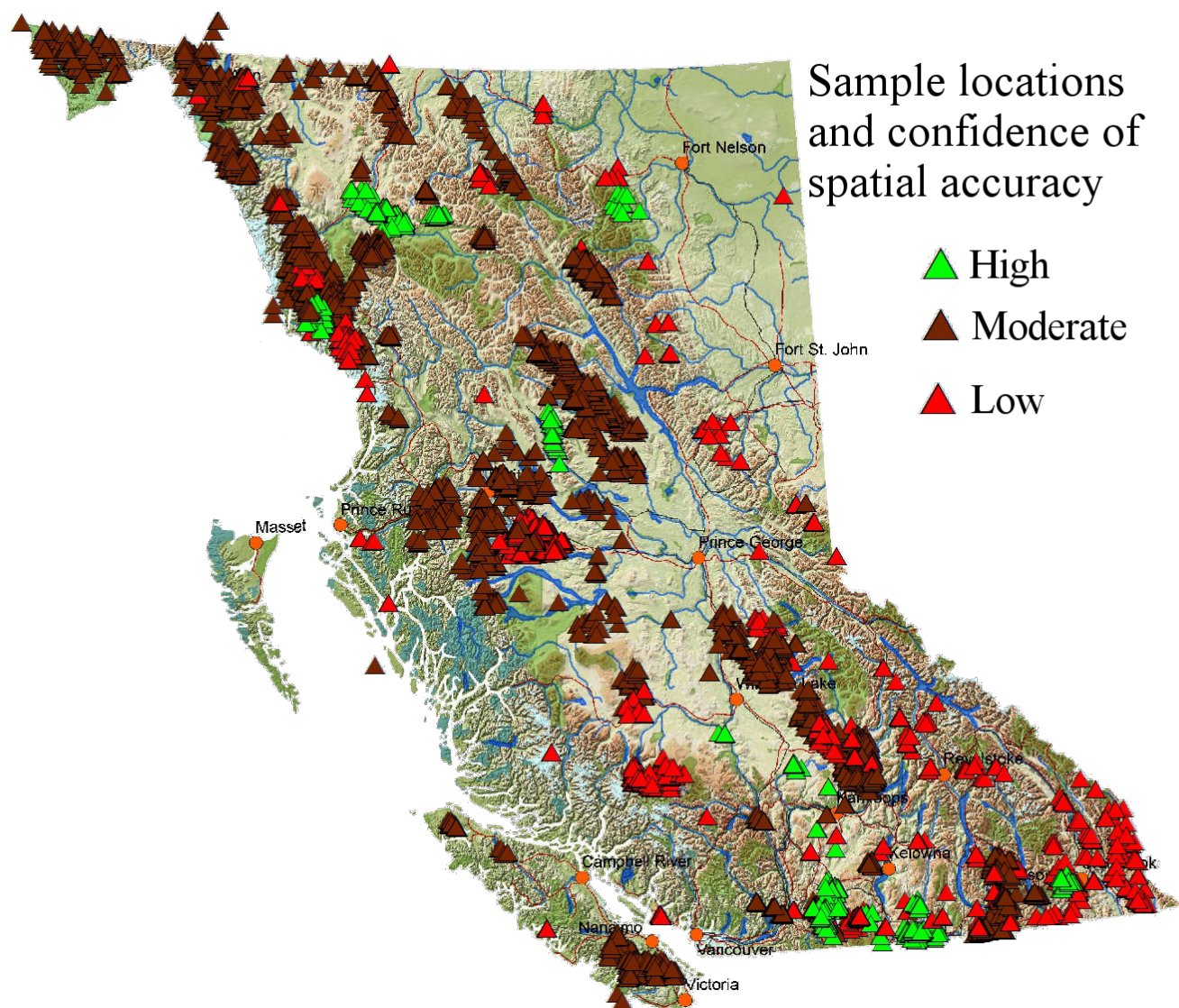
Fig. 3. Sample locations and confidence of spatial accuracy.

Analytical determinations below minimum Detection Limit (DL) are reported as negative values, consistent with other provincial geochemical datasets (Rukhlov and Naziri, 2015), so that users have a choice of handling these values as needed. Similarly, a value of '.99' is attached to determinations greater than the maximum DL. Consistent with the RGS database (Rukhlov and Naziri, 2015), missing values or nulls are reported as blanks instead of using an arbitrary value (e. g., '-1') to avoid conflicting with analytical results below minimum DL.

3. Database development

The commonly adopted database development cycle includes requirement analysis, logical design, implementation, and database population (Connolly and Begg, 1999). We followed this cycle closely to develop the British Columbia lithogeochemical database. We started by gathering feedback from the mineral exploration industry, examining the intrinsic relationships in the lithogeochemical data, and studying the

original database developed by Lett and Ronning (2005), which led to our logical data model (Fig. 4).

This model consists of 10 entities: **data_sample; code_ report; data_publish; data_analyte; code_detn; code_ geol; code_lab; code_method; code_prep;** and **code_unit**. The data_sample entity is for sample modeling and storing the sample-related information. The code_report entity is for publication modeling and storing publication-related information. The data_publish entity exists only to break up the many-to-many relationship between data_sample and code_report (which causes ambiguity in data query). The data_analyte entity is for analyte modeling and storing analyte-related information. The remaining entities are for modeling and storing the corresponding objects. They also function as lookup tables for maintaining data integrity in the database.

With minimum redundancy, this model can address relationships between sample, publication, analyte, analytical method, and laboratory, for example in cases where: 1) samples
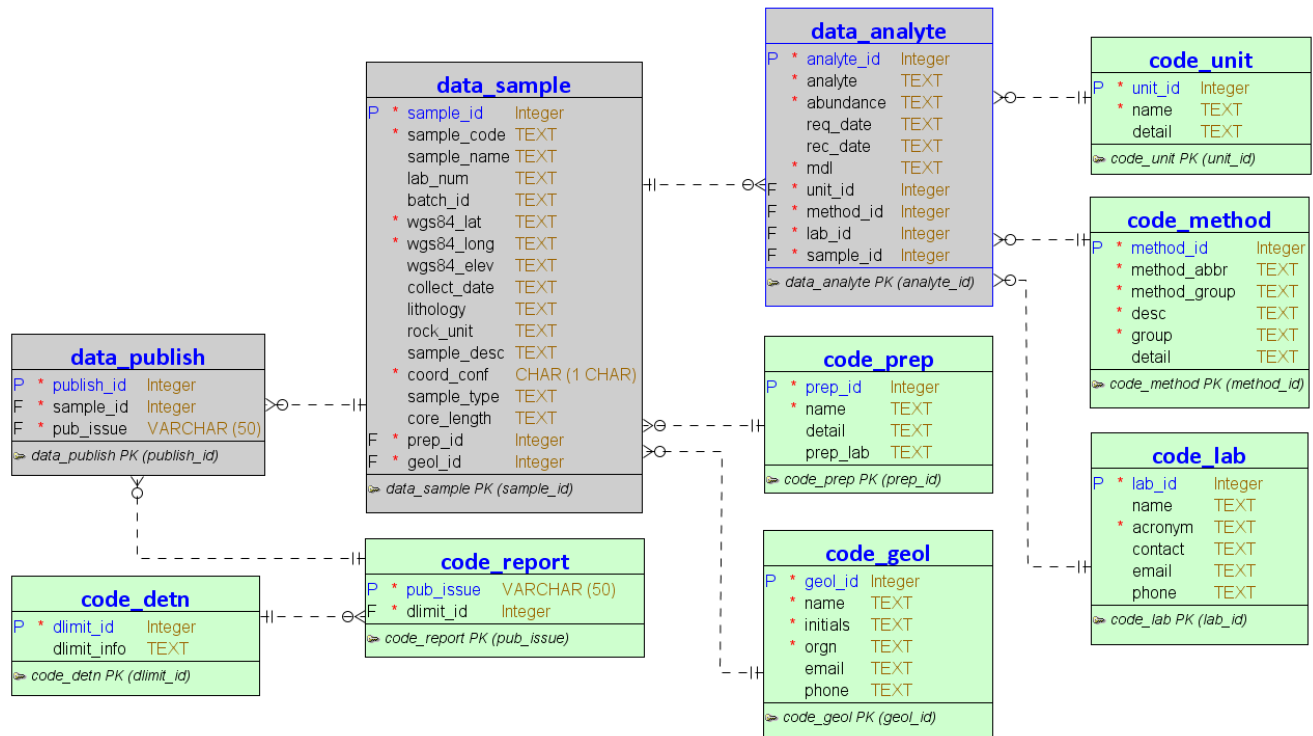
data_sample
P * sample_id — Integer
* sample_code — TEXT
sample_name — TEXT
lab_num — TEXT
batch_id — TEXT
* wgs84_lat — TEXT
* wgs84_long — TEXT
wgs84_elev — TEXT
collect_date — TEXT
lithology — TEXT
rock_unit — TEXT
sample_desc — TEXT
* coord_conf — CHAR (1 CHAR)
sample_type — TEXT
core_length — TEXT
F * prep_id — Integer
F * geol_id — Integer
data_sample PK (sample_id)

data_analyte
P * analyte_id — Integer
* analyte — TEXT
* abundance — TEXT
req_date — TEXT
rec_date — TEXT
* mdl — TEXT
* unit_id — Integer
F * method_id — Integer
F * lab_id — Integer
F * sample_id — Integer
data_analyte PK (analyte_id)

code_unit
P * unit_id — Integer
* name — TEXT
detail — TEXT
code_unit PK (unit_id)

code_method
P * method_id — Integer
* method_abbr — TEXT
* method_group — TEXT
* desc — TEXT
* group — TEXT
detail — TEXT
code_method PK (method_id)

code_prep
P * prep_id — Integer
* name — TEXT
detail — TEXT
prep_lab — TEXT
code_prep PK (prep_id)

code_lab
P lab_id — Integer
name — TEXT
* acronym — TEXT
contact — TEXT
email — TEXT
phone — TEXT
code_lab PK (lab_id)

code_geol
P * geol_id — Integer
* name — TEXT
* initials — TEXT
* orgn — TEXT
email — TEXT
phone — TEXT
code_geol PK (geol_id)

data_publish
P publish_id — Integer
F * sample_id — Integer
F * pub_issue — VARCHAR (50)
data_publish PK (publish_id)

code_detn
P * dlimit_id — Integer
dlimit_info — TEXT
code_detn PK (dlimit_id)

code_report
P * pub_issue — VARCHAR (50)
F * dlimit_id — Integer
code_report PK (pub_issue)

Fig. 4. Logical data model of the lithogeochemical database.

were analyzed using different methods at different laboratories; 2) samples were re-analyzed for a partial or full list of analytes; 3) samples were re-analyzed with different minimum detection limits; and 4) samples were published in multiple publications. With this model, we can unambiguously retrieve each determination with its associated information (sample, method, lab, and publication) using Sequential Query Language (SQL).

Applying the logical model, we constructed the new BC lithogeochemical database using Microsoft Access. Peripheral applications were developed as 4 Python scripts to assist routine data management activities, including data quality control, data loading, data extracting, and product generating. Our model aligns with the Open Geoscience data model (Granitto et al, 2012; Watson and Evans, 2012).

4. Database operation

The new lithogeochemical database makes it possible for the BCGS to centrally manage lithogeochemical data through time, which leads to improved data consistency, timely data update, routine delivery of customized data product, and data loss mitigation. The operation and dataflow of the database follows five steps (Fig. 5): 1) data compilation; 2) data QA/QC; 3) data loading; 4) product generating; and 5) data distributing.

In data compilation, information is located and retrieved from a variety of sources and saved to XLS files in a standardized format. Commonly, a single source fails to provide all the information required by the database, or multiple sources give inconsistent results. Data completing, verification, and correction are commonly needed to deal with these situations.

Data QA/QC is then conducted by the corresponding Python script that was designed to flag the most common data errors in the source or introduced in the data compilation step. These include errors such as unrealistic determinations and units, improper methods, incorrect sample locations, and redundant samples. The flagged errors are then corrected manually.

Following QA/QC, data are loaded into the database. This is an automatic step done by the corresponding Python script. To avoid generating duplicate entries, the script checks if a sample is already in the database.

Product generating, also conducted by a Python script, extracts data and outputs them in formats suited to the user (e.g., Comma-Separated-Value (CSV) files, ESRI shapefiles, or MS Access files). It is worth noting that database and the products derived from the database differ. First, the format of the products is based on data applications and can be changed by the users, whereas the format of the database is decided by the data model for database management and is fixed once database development is finished. Second, data products are commonly derived as the views of a database. As such, they only represent part of the information contained in the database. Separating the database from the product is important, because it helps the user generate multiple tailored products from a single database.

Distributing data by linking a database to a web-based GIS, such as MapPlace, the British Columbia Geological Survey map service, is a common approach. It can be achieved by executing Sequential Query Language (SQL) statements against the database constructed by assembling the user-provided input through the web interface.
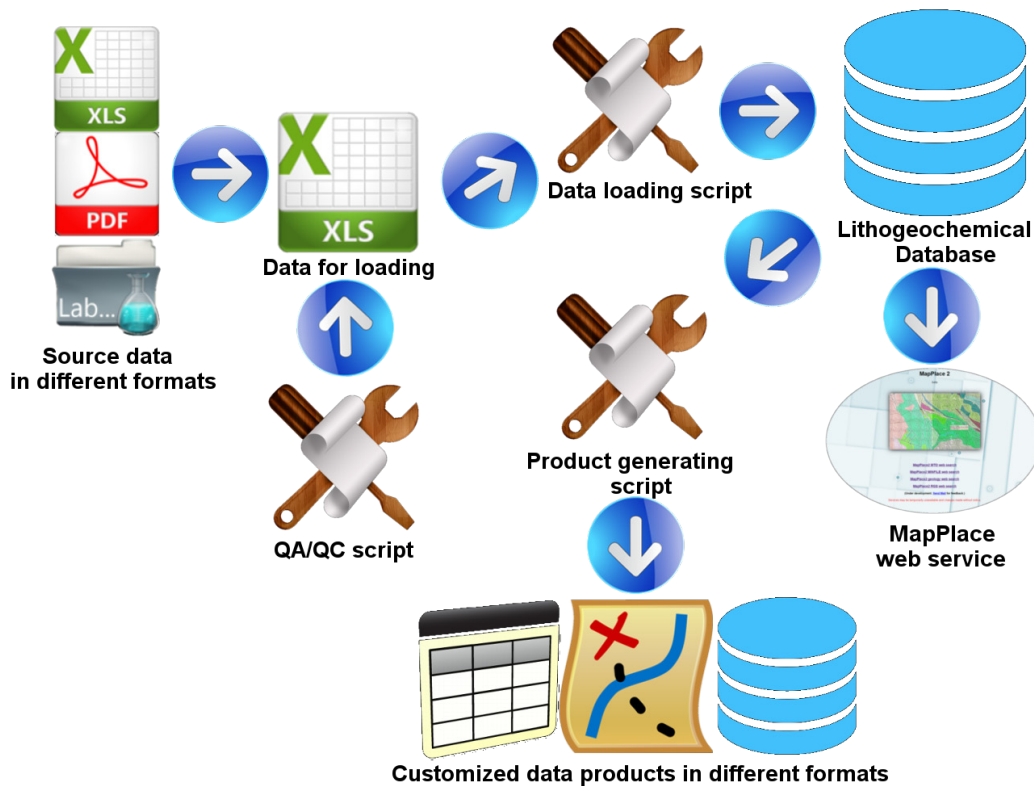
Fig. 5. British Columbia Geological Survey lithogeochemical data management.

5. GeoFile 2016-4 data product and format
This GeoFile includes three CSV files.
'lithogeochem_data.csv', the primary component of the release, contains sample-level information and analyte concentrations determined by various analytical methods (Table 1).

Table 1. Structure and content of 'lithogechem_data.csv'.

Attribute name	Description	Example
Sample_ID	Sequential number used as the row counter (generated by database)	5360
Sample_Name	Original sample or field names in publication or other data sources	CRE88-27
Sample_Type	Type of sample (surface or drill core)	surface
Sample_Desc	Sample description, which may include information about sample lithology, rock unit, mineralization, locality, or topography	Siliceous andesite
Pub_Issue	Identification of the publication where a sample was reported	P1992-04
WGS84_Lat	Sample latitude in WGS84	48.90037
WGS84_Long	Sample longitude in WGS84	-123.28742
WGS84_Elev	Sample elevation (m) above sea level	1250
Coord_Confidence	Confidence of sample location accuracy	H
BCGS_Grid	Mapsheet number in 1:20,000 BCGS map grid where a sample resides	92B.091
As_AAS_ppm	Arsenic concentration in ppm determined by Atomic Absorption Spectrometry (AAS), where samples were dissolved by aqua regia.	96
...	Concentration of next analyte determined by an analytical method	...

'lithogeochem_lab.csv' is a metadata file that provides information about detection limit, analytical laboratory, and sample pre-processing (Table 2).

Table 2. Structure and content of 'lithogechem_lab.csv'.

Attribute name	Description	Example
Data_Source	Publication identification	GF2001-08
DL_MAS	Representation of values below minimum detection limit (as appeared in the data source) for a given analytical method (in this example, MAS stands for multi-acid digestion atomic absorption spectrometry)	<DL
…	Representation of values below minimum detection limit (as appeared in the data source) for a given analytical method.	
Prep_Lab	Name of the analytical laboratory where sample preparation was conducted.	BCGS
Lab_XRF	Name of the analytical laboratory where analysis by a given analytical method was conducted (in this example, XRF means X-ray fluorescence).	
…	Name of the analytical laboratory where analysis by a given analytical method was performed.	

lithogeochem_method.csv is a reference table (e.g., Table 3) that provides details for all analytical methods and the codes adopted in lithogeochem_data.csv.

Table 3. Structure and content of 'lithogechem_method.csv'.

Attribute name	Description	Example
Method_Code	Abbreviation of an analytical method name	LE
Desc	Description of an analytical method	Leco combustion
Group	Partial or total determinations	Total determination
Detail	Analytical detail	Typically a 0.5g or 1g milled sample is mixed with a tin-copper conductor and heated in an oxygen stream in a closed system induction furnace. The SO_2 and CO_2 evolved are measured by an IR sensor or by volumetric/titrametric methods.

Acknowledgements

References cited

Connolly, T., and Begg, C., 1999. Database Sytems – A Practical Approach to Design, Implementation, and Management, 2nd Edition, Addison-Wesley, 113p.

Granitto, M., Schmidt, J. M., Labay, K. A., Shew, N. B., and Gamble, B. M., 2012. Alaska geochemical database - Mineral exploration tool for the 21st century. U.S. Geological Survey Open-File Report 2012-1060, 33p.

Lett, R., and Ronning, C., 2005. The BC rock geochemical database. British Columbia Ministry of Energy and Mines, British Columbia Geological Survey GeoFile 2005 - 14: http://www.em.gov.bc.ca/Mining/Geoscience/PublicationsCatalogue/GeoFiles/Pages/2005-14.aspx. Last accessed in Nov. 2016.

Rukhlov, A., and Naziri M., 2015. Regional Geochemical Survey Database. British Columbia Ministry of Energy and Mines, British Columbia Geological Survey GeoFile 2015-3: http://www.em.gov.bc.ca/Mining/Geoscience/PublicationsCatalogue/GeoFiles/Pages/GF2015-3.aspx. Last accessed in Nov., 2016.

Watson, C., and Evans, I., 2012. Geochemistry data model summary, British Geological Survey, http://www.bgs.ac.uk/services/dataModels/geochemistry.html. Last accessed in Nov., 2016.

British Columbia Geological Survey
Ministry of Energy and Mines
www.em.gov.bc.ca/geology