ESS / SST

Managing geochemical data - development of an appropriate database and delivery mechanism

Adcock, S.W., Laframboise, R.R., Spirito, W.A., Grunsky, E.C.

Environment and Health

North American Soil Geochemical Landscapes Project

Abstract

The management of geochemical data is inherently complex because of the multitude of sample media, preparation methods and analytical techniques. GSC staff have developed a universal model for the storage of geochemical data in relational databases. It is currently being used to manage a large amount of legacy data. The model has a very comprehensive set of tables for capturing metadata on all aspects of a geochemical survey (sample types, analytical procedures, etc.).

The data model provides a solid foundation on which to build computational procedures for managing the data. The procedures rely on a combination of SQL scripts and XML output. Fully automated ETL (extract, transform and load) procedures are used to convert data from legacy formats into a format that is easy to maintain, and then into formats that are easy to use. A sophisticated website comprising several thousand XHTML pages, KML files and spreadsheets is derived directly from the relational database (http://gdr.nrcan.gc.ca/geochem).

The management system is designed to be platform-independent, multilingual (currently English and French) and standards-based. It has been tested on several commercial RDBMSs. The system delivers CSDGM-compliant metadata to Z39.50 servers. A map-based query tool using University of Minnesota MapServer allows public geospatial searches.

The major development effort so far has focused on (a) methods to load legacy data into the database, and (b) transformation into web pages for end-users. We are now focusing on integrating the system with a laboratory information management system (LIMS). The existing LIMS will be redesigned to be more adaptable to the evolving needs of the organisation. It will be based on a logical n-tier (or n-layer) architecture. The system will be configurable for both internal GSC applications and external users via a web service.

Key features of the data management system

The foundation of the data management system is a generalized geochemical data model described in detail in a separate panel to the right. Minimal use of "in-house" software.

"In-house" software is highly modular, and minimizes dependence on specific vendors (avoids use of stored procedures in the database; favours use of "open" languages, such as SQL and XSLT).

Automated data loading procedures take input spreadsheet data and re-cast them into normalized tables.

Automated procedures re-cast the normalized database into a set of de-normalized tables, which can then be exported from the database in XML format.

XML schemas constrain the exported data.

XSLT procedures are used to transform the exported XML files into XHTML and KML files, suitable for posting to the Internet.

References

Adcock, S.W., Grunsky, E., Laframboise, R. and Spirito, W.A. (2003) A Universal Geochemical Survey Data Model. International Association for Mathematical Geological (IAMG) meeting, Portsmouth, UK, September 2003.

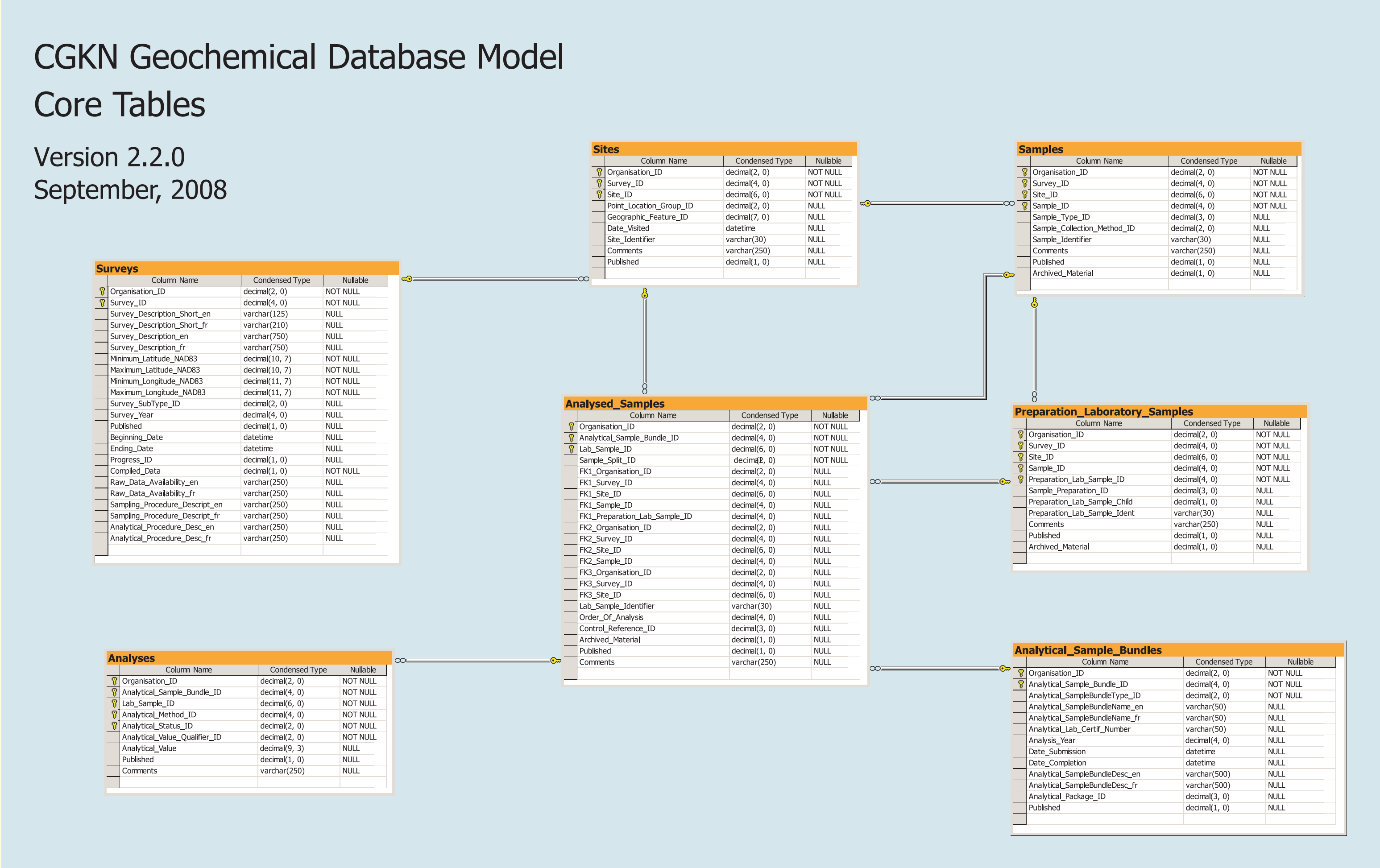
Halpin, T. (1995) Conceptual Schema and Relational Database Design. 2nd edition. Prentice Hall. 547pp

Acknowledgments

This work has been partially funded by Health Canada, under the Federal Contaminated Sites Action Plan.
A large part of the early development effort was supported by the Canadian Geoscience Knowledge Network (CGKN) initiative.
The authors thank Susan Davis for her graphical design expertise.

Contact Info:

Stephen W. Adcock: adcock@NRCan.gc.ca
Richard-Romeo Laframboise: rilafram@NRCan.gc.ca
Wendy A. Spirito: wspirito@NRCan.gc.ca
Eric C. Grunsky: egrunsky@NRCan.gc.ca



CGKN Geochemical Database Model
Core Tables
Version 2.2.0
September, 2008

Business process model for conducting a geochemical survey

(based on Adcock et al., 2003)

1. A principal investigator initiates a project.
2. As part of the project, a geochemical survey is carried out.
3. In the course of the survey, a number of sites are visited.
4. At each site, samples are collected.
5. Preparation laboratory samples are derived by physical processing of the material collected in the field.
6. Portions of the prepared material are separated out, and bundled together, to form analytical sample bundles. The bundles are then sent off to an analytical laboratory. Control Reference samples (for quality assurance) may be added to the sample bundles. Prepared material may be packaged into many different bundles and sent to many different laboratories.
7. The analytical laboratory performs the requested analyses.
8. The results of the project appear in various publications.

Laboratory Information Management System

In order for the GSC analytical laboratories to improve the delivery of essential services to the scientific community, the present Laboratory Information Management System (LIMS) will be redesigned. The service-oriented architecture of the new system will provide the flexibility to share some functionality with a wider public clientele through web service portals. The geochemical data model is an integral part of the system. Some features of the LIMS include sample registration and tracking, integration of QA/QC, traceability and automated report generation.

Support for metadata standards

The underlying data model contains fields that map directly to fields within the CSDGM v2 metadata standard.

A SQL Server 2000 implementation of the data model is used to expose metadata to Z39.50 servers.

A Z39.50 server is being maintained at the GSC, which is supplying metadata to Z39.50 search engines, including:
GeoConnections Discovery Portal:
http://geodiscover.cgdi.ca/gdp/search?action=entrySummary&entryType=productCollection&entryId=4133&entryLang=en

CGKN Data Catalogue: http://www.cgkn.net/cur/datacat/index.php
(select GSC - Canadian Geochemical Surveys)

Migration from CSDGM v2 to an XML-based (ISO 19139) standard will greatly simplify metadata maintenance procedures

CSDGM and ISO metadata standards

Content Standard for Digital Geospatial Metadata: Established by the Federal Geographic Data Committee (FGDC) in the USA in the 1990s, and subsequently adopted by Canadian government agencies.

ISO 19115 is an international standard for geospatial metadata, and ISO 19139 defines an XML-schema that is consistent with ISO 19115. Efforts are underway across North America to migrate from CSDGM to ISO 19115.
See http://www.fgdc.gov/metadata/geospatial-metadata-standards

Z39.50

A protocol for transmitting data cross the Internet (analogous to http and ftp). Developed in the Internet's infancy, it describes a way for one computer to ask another to do searches for it. It has been widely adopted by librarians, to allow them to search library catalogues.

Design features of the data model

Multilingual

Data types are a simple subset of those defined by XML Schema, which are easily translated to various RDBMSs.

Fully normalized, with many constraints to automatically ensure data integrity.

Clearly linked to a business process model for conducting a geochemical survey (see below).

Extensive collection of lookup tables which define the scientific terms that describe geochemical surveys.

Hierarchical classification schemes for (a) sample media, (b) decomposition techniques, (b) analytical techniques.

Straightforward correlation between the data model and the CSDGM metadata standard.

The model is designed to work with ongoing research programs, but will still work with legacy data where large amounts of metadata may be missing.

Online access to the data

Geochemical data stored in a normalized format in a large relational database are not easily accessible and usable by most geoscientists. Therefore we have developed procedures to automatically repackage the data into more convenient formats.

Raw analytical data have been packaged into MS Excel spreadsheets on a survey-by-survey basis. The spreadsheets have a standardized format and column naming convention. They are available in three different "flavours", corresponding to the three most popular methods of handling analytical values that are below the detection limit.

Analytical data are also available as KML files, with analytical values grouped by percentiles.

A simple web mapping application allows users to easily locate geochemical surveys geospatially.

[1] Analytical values below detection can be expressed as:
(a) Text values, using the less-than symbol (e.g. "<2")
(b) Numeric values, set to half the lower detection limit (e.g. "<2" is changed to "1")
(c) Numeric values, set to the negative value of the detection limit (e.g. "<2" is changed to "-2")

Alternatively, users can search for data for a particular element, using a periodic table query interface.

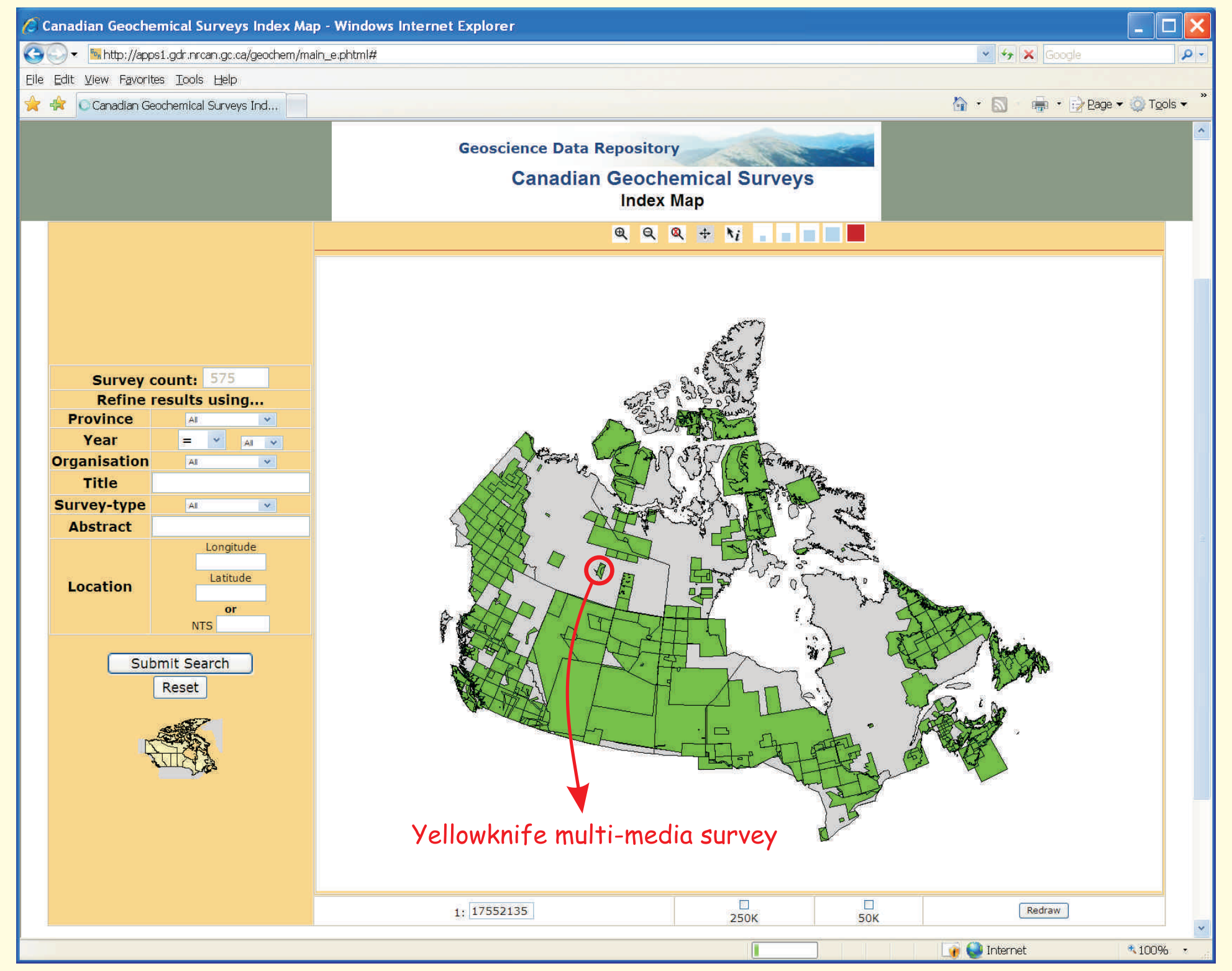
Historical development of the data model

The data model has its origins in numerous project-specific databases created by Adcock, Spirito, Laframboise and GSC colleagues in the 1980s and 1990s.
Based on their experiences, a generalized geochemical data model was created in the mid 1990s, using Halpin's (1995) object-role modeling methodology. The model was physically implemented in MS Access, SQL Server and Oracle. The model continues to evolve, but its core features (shown in the "core tables" schema diagram below) are essentially unchanged.

Adoption of the model across the GSC was slow because of (a) hardware and software constraints, and (b) the prohibitive cost of custom software development. Recent advances (most significantly in XML-based technologies) have mitigated these factors, and the model has been used for several recent projects.

Geospatial searching for geochemical datasets

Geospatial searches of the metadata catalogue are based on University of Minnesota MapServer software. The screenshot shows outlines of the 575 surveys catalogued to date. Surveys carried out in Canada in 2007 as part of the North American Soil Geochemical Landscapes Project are currently being catalogued.

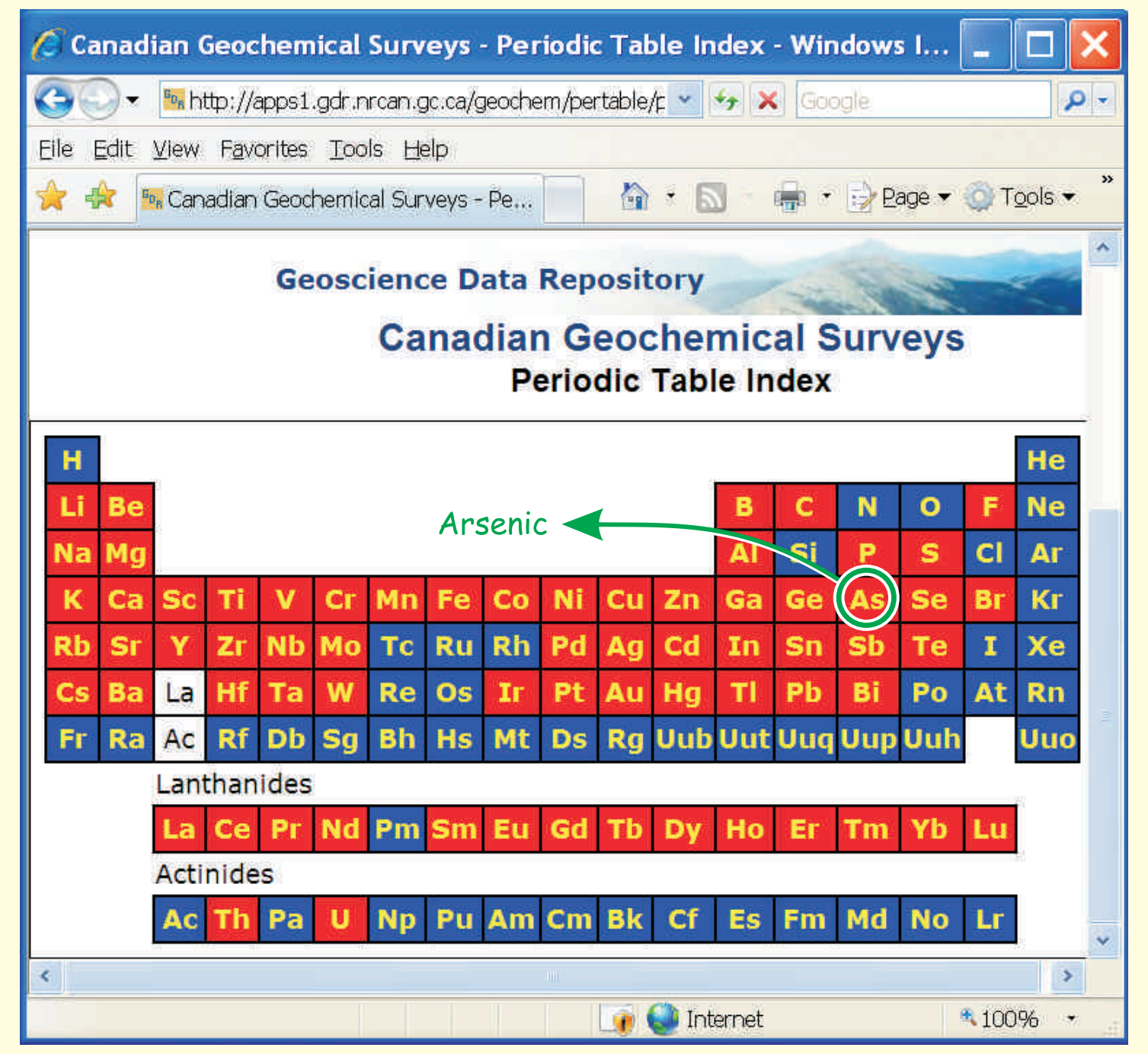


Canadian Geochemical Surveys Index Map - Windows Internet Explorer
http://apps1.gdr.nrcan.gc.ca/geochem/main_e.phtml#

Geoscience Data Repository
Canadian Geochemical Surveys
Index Map
Yellowknife multi-media survey

Periodic Table Query Interface

This screenshot identifies those elements for which geochemical data are available online in standardized formats (MS Excel spreadsheets and KML files). Each of the red boxes is hyperlinked to a KML file containing an index of surveys which include analytical data for the specified element.



Canadian Geochemical Surveys - Periodic Table Index - Windows I...
http://apps1.gdr.nrcan.gc.ca/geochem/pertable/

Geoscience Data Repository
Canadian Geochemical Surveys
Periodic Table Index

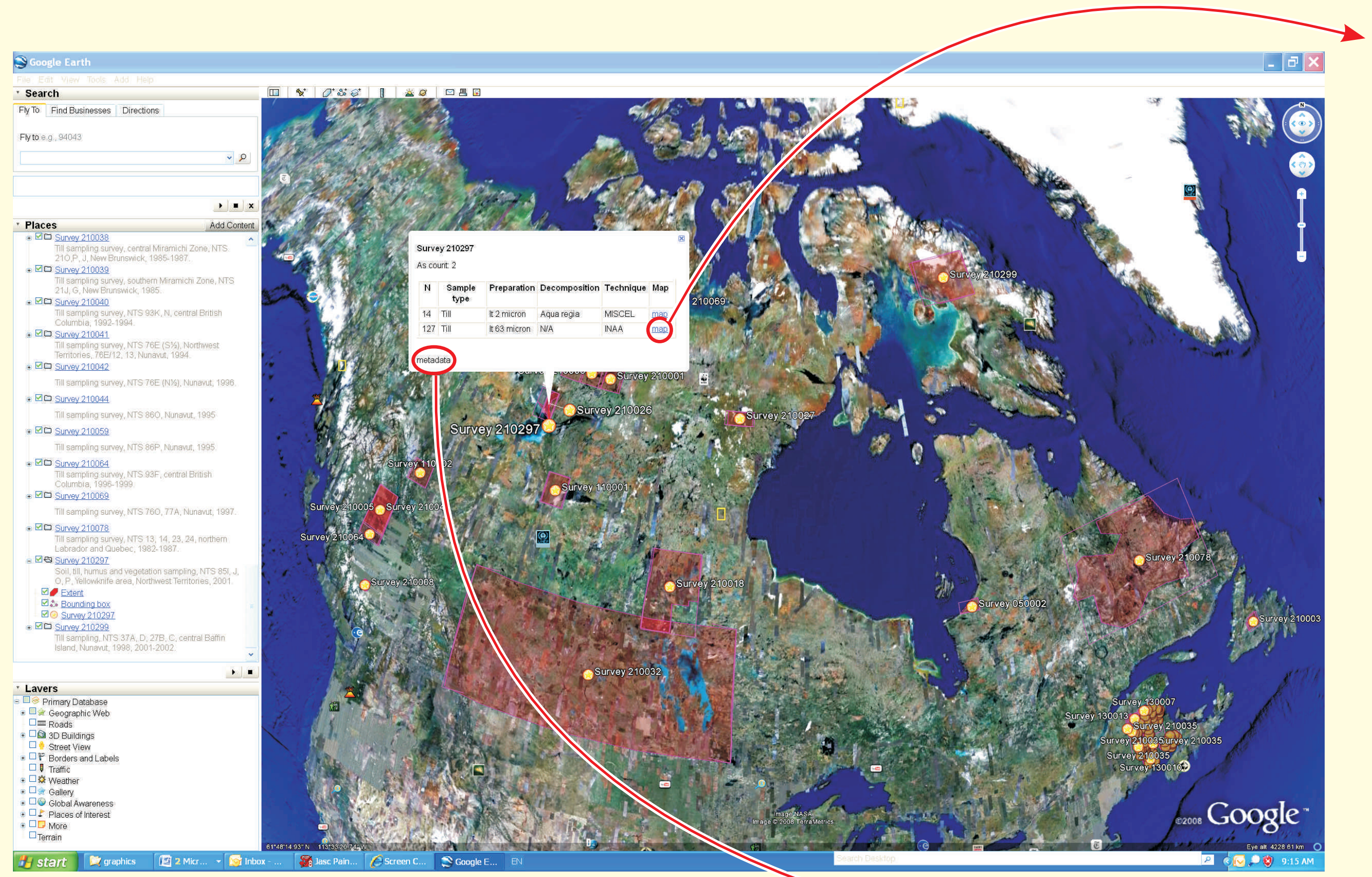
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba La Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Ac Rf Db Sg Bh Hs Mt Ds Rg Uub Uut Uuq Uup Uuh Uuo

Lanthanides
La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu

Actinides
Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr

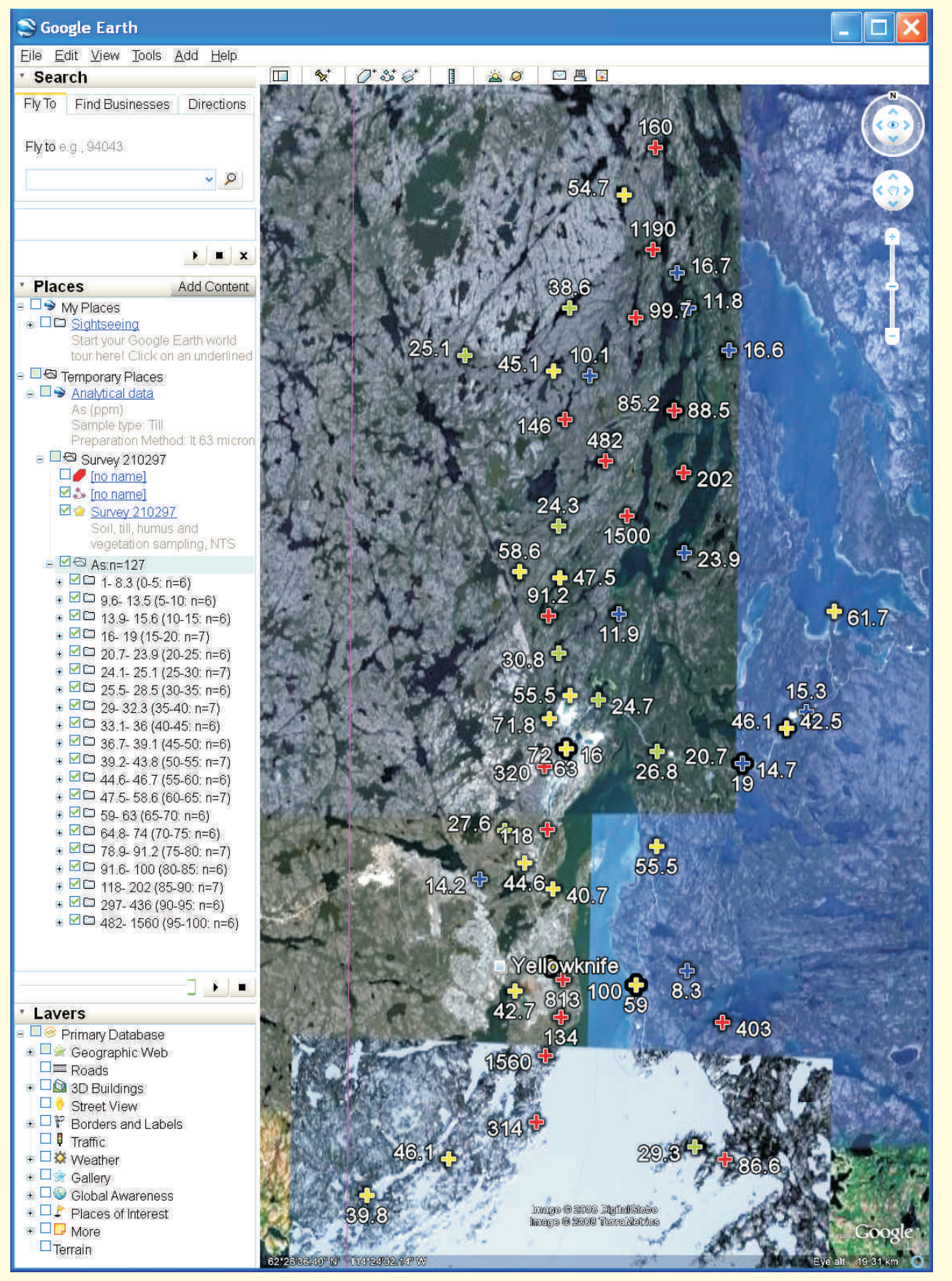
KML file showing surveys for which arsenic data are available online

This screenshot shows the index map for surveys that include arsenic (As) data.



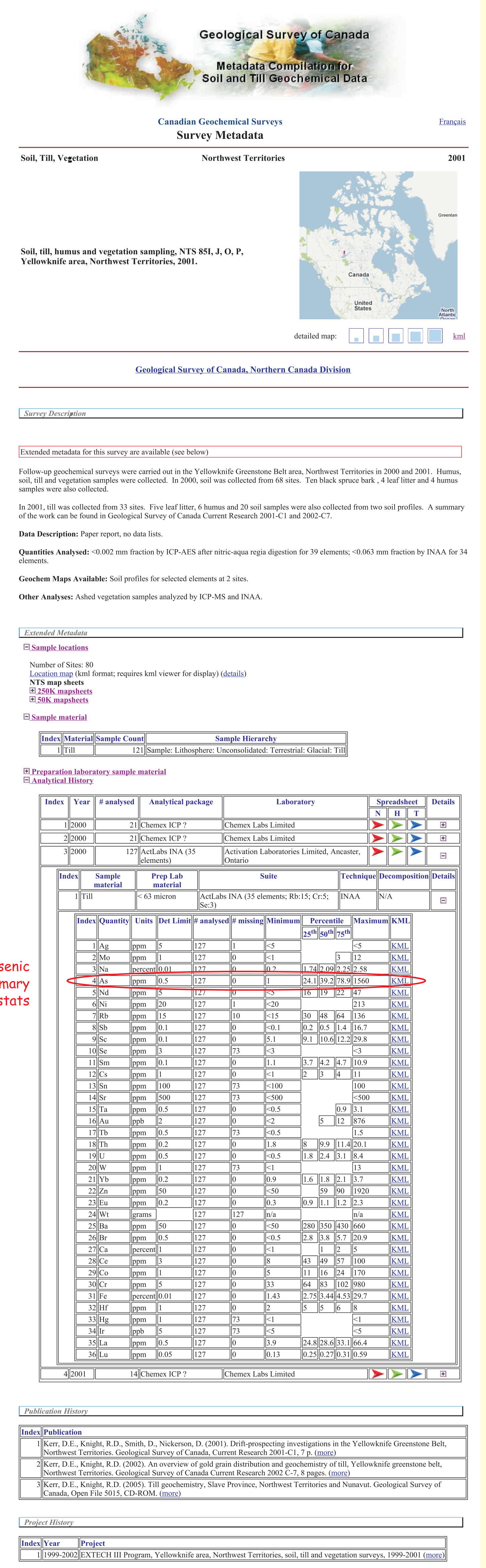
KML file showing arsenic analytical values

This map shows arsenic INAA data from the < 63 µm size fraction of till samples collected around Yellowknife, NWT. The analytical values are colour-coded by quartiles, and the legend breaks the data down by 5-iles.



Web page showing standardized presentation of metadata for a geochemical survey

This example page describes a multi-media survey around Yellowknife.



Arsenic summary stats

Natural Resources Canada
Ressources naturelles Canada
Canada