



Final Report to the Canadian Geoscience Data Model Working Group

Recommendations for a Geoscience Model to
Support an Integrated, Internet-based Geoscience
Knowledge Network for Canada

June 2000

Prepared by:

Marla Weston, Ph.D

Henry Kucera, M.Sc, P.Geo

Gail Kucera, Ph.D

Executive Summary

This is the final report of a project initiated on behalf of the Canadian GeoScience Knowledge Network (CGKN) to collect information on current geoscience data management standards and models in Canada. The project used three investigative methods.

- Reviewed standard models in current use within various geoscience sectors.
- Surveyed the 18 geoscience agencies participating in the CGKN.
- Participated in a workshop on geoscience data models in Calgary on June 4-5 to discuss requirements and technical alternatives.

This report summarizes the results of the investigation and provides recommendations on how to proceed with developing a "CGKN Webhouse"—a single web portal by which to access geoscience data Canada-wide. The purpose of this document is to bring together the findings from all the components of the Geoscience Data Model Review project, present an analysis, and recommend an approach to development.

The review of standard models found that no single model encompasses all the geoscientific requirements. However, each standard model has elements that could be adopted for the CGKN common model. Adhering to the approaches of the standard models is highly desirable because of the work that has gone into each, and the fact that each standard model presently serves a sector of the intended CGKN audience.

The survey of 18 geoscience agencies found that there is no widespread consensus on the best model, and that no standard model is in widespread use. Some agencies have invested considerable funds into developing database systems that meet their corporate needs; those agencies will be unable to change their core data model without a period of transition. Some agencies are only beginning to implement automated systems. These agencies could benefit from the experience of others, but only if it is offered in a cost-effective way.

There was consensus among the surveyed agencies that the best first step toward an integrated, internet-based means of sharing geoscience data across Canada was to develop a standards-based, web-enabled catalogue of geoscience data across Canada. The catalogue process would add standard metacontent¹ to all data sets, and register each available dataset by recording its metacontent in an on-line geoscience data registry. Each agency would be responsible for providing the metacontent for their data holdings.

The geoscience data registry would be accessible via the CGKN website—a single portal by which to access geoscience data Canada-wide. This portal would allow the

¹ Metacontent is data that describes the subject, lineage, points of contact, structure, and other characteristics of data.

user to search for geoscience data by geographic area or by subject. Using the search criteria, the portal would direct the user to the web site of the data's custodian. Development of this capability is "Component 1" of the recommended approach.

The recommended approach involves four components of work. The components deliver in stages. The goal is to provide some capability as early as possible by undertaking straightforward tasks first (i.e., the geoscience data registry). Early delivery also is possible by leveraging existing work. The four recommended components of work are as follows.

Component 1	Component 2	Component 3	Component 4
<ul style="list-style-type: none"> • Introduce and populate minimal meta-content model • Single CGKN portal. • Spatially enabled search • Services registry that links to agency web sites. • Common geographies used to navigate to information assets. 	<ul style="list-style-type: none"> • Enhance and populate extended meta-content model • Browse externally held data assets through single portal. • First iteration of the common model for a limited set of subject areas (e.g., bedrock, geology, soils, etc.) • Integrate recent technical improvements. 	<ul style="list-style-type: none"> • Develop common logical geoscience model. • Collaborate with other interested parties. • Extend the common model to include a wider set of subject categories with feature class definitions. 	<ul style="list-style-type: none"> • Implement common model • Assist agencies that are willing to adopt model elements.

A complete project plan has not been developed as part of this investigation. However, it is possible to suggest an appropriate timeline for each component. The timelines assume that components can be constrained to meet the suggested schedule in the interest of having early and progressive deliveries.

ID	Task Name	Start Date	End Date	Duration	2000		2001			
					Q3	Q4	Q1	Q2	Q3	Q4
1	Develop Project Plan	7/1/00	7/21/00	15d	15d					
2	Initial CGKN Portal	8/21/00	12/21/00	89d	89d					
3	Improved CGKN Portal	10/15/00	8/16/01	219d	219d					
4	1st Iteration Data Model	8/21/00	1/19/01	110d	110d					
5	Data Model Implementation & Application Port	1/22/01	10/22/01	196d	196d					

Table of Contents

Executive Summary..... i

1 Introduction..... 1

1.1 Purpose of This Report.....2

1.2 Organization of This Report.....2

2 Review of Candidate Models..... 3

3 Summary of Survey Results 5

3.1 Metadata and Data Format Standards.....5

3.2 Data Models.....6

3.2.1 Geological Maps..... 6

3.2.2 Minerals/Hydrocarbon Databases..... 7

3.2.3 Geochemistry/Geophysics Databases (Raster and Vector)..... 7

3.2.4 Assessment Reports and Associated Databases..... 7

3.2.5 Borehole Databases..... 8

3.2.6 Biostratigraphic Databases..... 8

4 Major Issues 8

5 Review of the Business Drivers 10

5.1 Intra-Sector and Cross-Sectoral Connectivity 10

5.2 Inventory of Available Data 11

5.3 Manage and Distribute Data 12

5.4 Synthesize and Analyze Information..... 12

5.5 Support for Secondary or Value-Added Usage..... 13

5.6 Support for Non-traditional Information Applications..... 13

5.7 Improving Overall Efficiency 14

6 Discussion of Metacontent 15

7 Recommendations..... 18

7.1 A Multi-Component Approach for Development.....20

7.2 Component 1: CGKN Portal and Meta-information Model 22

7.2.1 The Extensible Metacontent Model 23

7.2.2 Details on the Repository Browser and Services Registry..... 27

7.3 Component 2: Integrated Spatial/Subject Web Client.....32

7.4 Component 3: Develop a Common Data Model.....36

7.5 Component 4: Implement the Common Data Model in Stages38

8 Conclusion 38

9 Appendices 41

1 Introduction

Geoscience data has a wide audience with diverse needs. Its users want to be able to access geoscience data directly, at their convenience, and in a form that is familiar and easy for them to use. Decision-makers want to generate reports on their own workstations or thin clients. A geoscientist may want access to a variety of background or historical information, as part of a study. A prospector may want a colour plot of claims and mineral occurrences. If a distributed spatial data network is properly configured, users should be able to issue ad hoc queries to explore trends, identify problems, evaluate market opportunities, or order data to respond to application-specific requirements. This is already possible in organizations that have dedicated subject-specific implementations; but it is not yet possible to share data freely across the GSC or between the members of the broader geoscientific community in Canada. To meet the demand for an integrated geoscience data source, organizations will need to find commonality in the fundamental ways they describe their data.

The Canadian Geoscience Knowledge Network (CGKN) is an Internet-based network for access to geoscience information from Canada's federal, provincial, and territorial geological surveys, universities, and other partners. The CGKN was created in 1998 by the National Geological Surveys Committee. The CGKN's goal is to improve access to information about Canada's geoscience resources, so it can be used more easily by the mineral and energy industry, public policy and regulation decision-makers, educational institutions, and all Canadians. The Canadian Geoscience Data Model Working Group (CGDMWG) is defining the technical means by which the information should be communicated.

This is the final report of a project initiated by the CGDMWG to collect information on current geoscience data management standards and models in Canada. The project used three investigative methods.

- Reviewed standard models in current use within various geoscience sectors.²
- Surveyed the 18 geoscience agencies participating in the CGKN.³
- Participated in a workshop on geoscience data models in Calgary on June 4-5 to discuss requirements and technical alternatives.

The survey gathered information on the use of public and standards-based models, and other models (or data structures) put forward by the stakeholders. The survey also gathered comments and recommendations from the various stakeholder agencies about how they would like to proceed toward a common infrastructure for information

² See "Evaluation of Candidate Models for GeoScientific Data," March 31, 2000, by Holonics Data Management Group Ltd.

³ See "Survey of Geoscience Agency Systems and Data Models," May 26, 2000, by Holonics Data Management Group Ltd.

access. The survey results indicated that the best first step toward an integrated knowledge network is to develop a geoscience meta-information model that includes a subject classification and other information components.

The meta-information model would support a common web portal to distributed Canadian geoscience data. This portal would allow the user to search by geographic area or by subject. Using the search criteria, the portal would direct the user to the appropriate web site. For some agencies, creating even a simplified form of metadata will be a significant undertaking; however, all of the surveyed agencies wished to participate in this process.

1.1 Purpose of This Report

This report summarizes the results of the investigation and provides recommendations on how to proceed with a data model and technical solution to further the data-sharing goals of the CGKN. The purpose of this document is to bring together the findings from all the components of the Geoscience Data Model Review project, present an analysis, and recommend a development approach to be undertaken by the CGDMWG.

1.2 Organization of This Report

This document is organized as follows.

- Section 2 summarizes the evaluation of candidate models for geoscientific data, and clarifies some of the issues related to these models and their ongoing development.
- Section 3 summarizes the results of the survey regarding the use of corporate data models and standards by Canada's geoscience agencies.
- Section 4 highlights the major issues and challenges that were discussed in the previous two detailed reports.
- Section 5 outlines the business drivers that must be supported by a CGKN "Webhouse" ⁴ and recommends a strategy for a balanced development approach that would provide the maximum benefit to the members of the CGKN.
- Section 6 Discusses the different uses of metacontent and describes how the CGKN could best leverage it.
- Section 7 provides a tactical plan that proposes the development of four technical components to meet the requirements of the CGKN.

⁴ Kimball, R., (1999), "The Webhouse has no Center." *Intelligent Enterprise*, July 13, Vol. 2, No. 10, Miller Freeman. "Every data warehouse must now be made available through Web browser interfaces..... And the data warehouse is being asked to make the customer clickstream available for analysis. These forces are changing the way we design and implement the data warehouse. As a signal of these changes, I have renamed my column **Data Webhouse**..... the data webhouse has no center, because it is unavoidably distributed. The data webhouse is not only distributed within individual organizations; it is also distributed among multiple organizations."

2 Review of Candidate Models

An earlier report⁵ reviewed four public geoscience data models: NADM v5.2 (CORDLink), PPDM (version 3.5 beta), POSC (Epicentre v2.2), and ODP (JANUS). This section summarizes the findings of that report.

Each of the four models has strengths and weaknesses as a candidate for the Canadian Geoscience Data Model. None of the four models provided a complete "turnkey common model" to integrate data holdings and information delivery.

Both PPDM and POSC Epicentre were developed by and for the petroleum industry. The Public Petroleum Data Model (PPDM) was developed by an association that represents over 100 oil and gas companies, vendors, and regulatory agencies worldwide. Their mandate is to deliver "...a vendor-independent standard petroleum data model that serves as the industry foundation for managing information as an essential asset in the global business of oil and gas exploration and production" (<http://www.ppdm.org>, updated May 18, 2000). PPDM v3.4 covers the subjects that are shown in Figure 2-1.

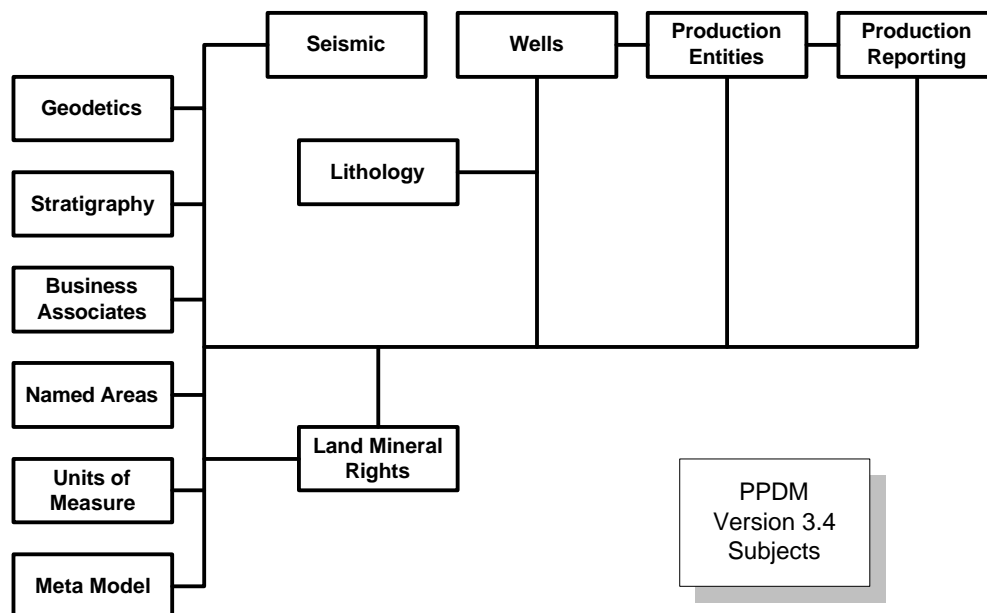


Figure 2-1: Subject Classification Scheme for PPDM v3.4.

Although PPDM has extensive breadth and depth, it is focussed on the needs of the oil and gas industry. PPDM is in the process of changing to become spatially enabled and to allow a more flexible approach to stratigraphy, among other areas.

⁵ See "Evaluation of Candidate Models for GeoScientific Data," March 31, 2000, by Holonics Data Management Group Ltd.

POSC Epicentre is the most fully specified of all the models reviewed, but it is also the most complex. It would therefore be difficult and expensive to implement. In addition, Epicentre in its pure form is strictly a logical data model, and as such, it is not directly implementable as a physical database. The Petrotechnical Open Software Corporation offers consulting to help implement the model, but this is likely to be a costly alternative.

The ODP JANUS data model is more science-based than either POSC or PPDM. One of its greatest strengths is a clear separation between interpretive attributes, and source attributes or measured attributes. It also can link interpretive data to a scientist or a published source. This is particularly important if data are collected in an environment that is not dominated by a single set of standards. JANUS also includes paleontological, geochemical, and geophysical specifications, along with extensive borehole coverage. The JANUS model does not, however, specifically include geological map data to the same extent as CORDLink.

NADM v5.2 (CORDLink) was reviewed as part of this study, and was described in the report that evaluated candidate data models. After that report was completed, it was pointed out that the use of the term "North American Data Model" is restricted to NADM v4.3. The development of the NADM v4.3 data model was coordinated by the AASG/USGS Geologic Map Data Model Working Group, the state geological surveys (of the USA), and the Geological Survey of Canada. The stated purpose of NADM v4.3 is "...to provide a structure for the organization, storage, and use of geologic map data in a computer."⁶ As a result, this model focuses almost strictly on geological maps and immediately supporting entities.

CORDLink (also known as NADM v5.2) is an extension to NADM v4.3. The CORDLink vision is to provide a comprehensive, flexible resource for scientific and non-scientific users of geoscience information; thus, the CORDLink vision extends beyond geological maps. Another key difference between CORDLink and the NADM v4.3 model is the move away from implementation-specific data structures. CORDLink was intended to be a logical model at a higher level of abstraction than the NADM v4.3 model. This goal has been partially achieved, leaving some issues to be addressed and enhancements to be made to complete the model

It is the CORDLink model that was reviewed in the "Evaluation of Candidate models for Geoscientific Data" report, since CORDLink incorporates a broader spectrum of the geosciences. To minimize the confusion between NADM v4.3 and CORDLink (NADM v5.2), hereafter the name "North American Data Model" will be used only when referring to NADM v4.3 and the name "CORDLink" will be used to describe NADM v5.2.

Like the ODP JANUS model, the CORDLink model is science rather than industry based. CORDLink maintains its focus on geological maps and includes interpretive data. The vector representation within CORDLink is especially well documented, and the geoscientific model approach is an essential element. However, development of the

⁶ B.R. Johnson, B. Brodaric, G.L. Raines, J.T. Hastings, and R. Wahl (1999) DIGITAL GEOLOGIC MAP DATA MODEL Version 4.3. Available at: <http://geology.usgs.gov/dm/>

CORDLink model is not complete, and parts of the model are poorly documented or limited in scope. This makes it difficult to determine if CORDLink will be effective for such disparate disciplines as paleontology, geochemistry, and geophysics, among others.

As noted in the original report, the Canadian Geoscience Data Model must provide a balance between an industry focus and a scientific focus. None of the four candidate models accomplishes this, although the complex POSC Epicentre provides a framework that comes closer than the other three. Each data model has its strengths and weaknesses. The Holonics evaluation report suggested that a combination of elements from each of the four models could provide an initial core for the development of a complete Canadian Geoscience Data Model. This suggestion remains valid, with the caveat that it also will be important to factor in the modelling approaches of the 18 territorial, provincial, and federal geoscience agencies.

3 Summary of Survey Results

This section summarizes the data model survey, including the interviews conducted with the 18 territorial, provincial and federal geoscience agencies.⁷ It also provides a more focussed evaluation of the data models used by each agency.

3.1 Metadata and Data Format Standards

Among the agencies reporting that they use or plan to use a metadata standard, over 80% listed the Federal Geographic Data Committee (FGDC) standard for digital geospatial metadata as their preference. In some cases (e.g., Ontario and Newfoundland) a version or distillation of the FGDC standard was used, but the underlying features were FGDC-based. GILS, ISO TC 211, OpenGIS and SDTS also were reported to be used for a few systems.

Internationally recognized data format standards specific to a data type appear to be relatively uncommon except for geophysical data sets. SEED (for natural source seismic data), SEG-Y (for other seismic data), and SINEX and RINEX (for GPS data) were mentioned. In Alberta, LAS and Synthetic LAS data formats also are used. Those agencies using an international data format standard emphasized the importance of supporting the standard. Some of the agencies stated that it was more important for them to conform to the defined data format standard than adopt another system, even if it was proposed as part of a common Canadian geoscience model.

Many provincial agencies have their own, internal standards. Some of the standards are very well developed and complex, and have a long history of use.

⁷ For a complete discussion, see "Survey of Geoscience Agency Systems and Data Models," May 26, 2000, by Holonics Data Management Group Ltd. The report includes the survey forms as completed by the survey respondents.

3.2 Data Models

A total of 106 individual data sets were reported by participating agencies for the following categories: geological maps (topology and attributes), minerals and hydrocarbons, geochemistry and geophysics, assessment reports and associated databases, boreholes, and biostratigraphy. Of these:

- only 10 were reported as relying on a public data model,
- 6 of 10 used PPDM,
- 2 of 10 systems were based on either NADM v4.3 or CORDLINK,
- 1 of 10 used the ODP JANUS model in part, and
- 1 of 10 partially used the British Georecords model.

It is interesting to note that less than 10% of systems or data sets reported in the survey used a non-proprietary data model. This probably points to one of two problems: either a non-proprietary model is not available for that particular type of data, or the non-proprietary model is too difficult to implement. Many agencies indicated that even if a data model were available, they did not have the time, resources, or expertise to develop or maintain a system based on that model. Thus, there is a requirement for education and outreach to ensure effective use of the model as the underpinning of the CGKN.

3.2.1 Geological Maps

Although North American Data Model (NADM v4.3) and the CORDLINK model were used to build only two of the 24 systems reported in this category, there is significant interest in these models. Many agencies either have plans to use one of the models in the near future or are waiting for more evidence of their viability before implementing a system using one of them.

The Ontario Geological Survey plans to implement the geological mapping segment of their system in a combination of the CORDLINK model and Land Information Ontario (LIO). The Ontario Geological Survey and Ontario as a whole are looking to LIO as an environment to deliver all their products online. It is a relatively new initiative of the Ontario Ministry of Natural Resources and will likely mutate to incorporate all the "knowledge" of challenges faced and resolutions found in the complex world of web delivery.

Geological maps appear to be most strongly influenced by the software used to visualize them and publish them to the web. Even agencies with a firm commitment to the use of the CORDLINK model have had to modify their implementations to meet the needs of the software. This will almost certainly improve as newer software and techniques are used for publishing to the web.

3.2.2 Minerals/Hydrocarbon Databases

No data model standard was mentioned for any of the 14 data sets reported in the minerals category. However, there are some large and mature systems reported for handling mineral occurrences. Among these are MINFILE (BC), NORMIN.DB (NWT), and MODS (NF). Both Ontario and Quebec have extensive coverage of mineral deposits in their systems, and other agencies have systems that deal with mineral rights and coal.

Eight data sets were recorded under the hydrocarbon category and half of these (4) were reported as being based on PPDM. The breadth and depth of the PPDM model, along with its industry support, makes it a very popular model for agencies dealing with hydrocarbons and related data types.

3.2.3 Geochemistry/Geophysics Databases (Raster and Vector)

No geochemical data models were reported for any of the 11 entries in the geochemical database summary table. Manitoba, Ontario, Quebec, and New Brunswick all have large geochemical databases based mainly in Ingres or Oracle.

Although not included in the geochemical summary table, the Terrain Sciences Division, in cooperation with the Mineral Resources Division and the Continental Geosciences Division of the GSC in Ottawa, has developed a geochemical data model called the "MultiDivisional Database Model." A test website has already been completed based on this model, and a new project called "GeoChemistry Online" will seek to complete and enhance the single-access window to National Geochemical information. Three provinces--Manitoba, Saskatchewan and Nova Scotia—also are involved in the GeoChemistry Online initiative. Given the cooperative approach to the development of this model, it will be important to review it upon its completion as a basis for a Canadian standard.

Only one data model, PPDM, was noted for the 21 data sets reported for geophysical data. PPDM supports seismic and other hydrocarbon/well based geophysics. For aeromagnetic and gravity data, the National Aeromagnetic Database and the National Gravity Database maintained by the GSC in Ottawa are large, mature systems.

3.2.4 Assessment Reports and Associated Databases

No public, standard data model was reported for any of the eight assessment report databases listed by participating agencies. However, it is worth noting that ARIS (BC) already offers web-enabled assessment reports. Manitoba, Ontario, New Brunswick, and Newfoundland all have extensive systems that cover assessment reports.

3.2.5 Borehole Databases

Although no public data models were listed for the 12 borehole databases, PPDM would be an option with ODP JANUS as an alternative for non-production boreholes.

3.2.6 Biostratigraphic Databases

One of the five entries for biostratigraphy lists PPDM as its data model standard, and another entry notes that PPDM is expected to be used the future. Significant changes have been proposed for PPDM v3.5 to support greater stratigraphic functionality. These changes include allowing the user to create more than one set of stratigraphic names. This will likely make PPDM more attractive for systems in this category.

The only other data model mentioned is the British Georecords model. This model was developed by the Cambridge Arctic Shelf Programme (CASP) at the University of Cambridge over 15 years ago. It is no longer being supported.

Four paleontology databases were included. All of these are within the GSC but appear to be independent of each other, and do not use a common data model. The ODP JANUS model does support paleontology, but none of the agencies with paleontology databases have applied this model.

4 Major Issues

The geoscience agency survey raised a number of issues that should be addressed to ensure that a comprehensive model or internet-based system is fully effective. Key issues are as listed in Table 4-1.

Table 4-1: Issues to be addressed for effective development.

Issue	Ramifications for improved data access
No single data model is being used, or is felt to suffice.	A pluralistic approach to data modelling is needed for the near future.
No single approach for metadata or data management is being used.	Many challenges exist for searching data holdings and exchanging data.
A lack of funding makes discretionary spending difficult. Many agencies would like long-term funding in return for participating in federal initiatives.	Some agencies are unable to participate in GSC initiatives to improve or standardize data access because they do not have the funding, or because they are unwilling to commit their own funding without a similar show of commitment by GSC.
Cost-recovery policies exist for geomatics data from Geomatics Canada, as do local user fees for data.	Any initiative to improve access to data will be made more complex by the patchwork of policies governing cost recovery. The ultimate costs recovered are arguably offset by the cost of cost recovery.

Issue	Ramifications for improved data access
Multilingual requirements are unclear.	The patchwork of policies and practices create a challenge, particularly if federal language laws are strictly interpreted for local scientific data. For some aboriginal languages, no translation is possible.
Need for client input to the modelling process.	There is a diverse clientele for geoscience products. Good representation from industry is required if modelling standards are to be established.
In the past, GSC has started, then dropped promising developments.	Many agencies feel unwilling to commit their own resources to developments that can be dropped without consultation.
Leaders in web development lose out on "catch up" funding.	There is a sense that those who spend local monies for technical advancement are penalized when later funding is expended to help others catch up. At minimum, the lessons learned by the early innovators should be made available to others as they undertake similar developments.
Giving credit where credit is due, with some balance.	If territorial and provincial agencies accept even minimal GSC participation on a project, they also must accept three GSC logos on the results. The three logos eclipse the single local logo, and make requests for local funding more difficult to justify.

The results of the survey indicate that few agencies (<10%) are using any public data model. As previously noted, there are two possible causes for such a low percentage: no public data model exists for that type of data, or the public data model is too complex, difficult, or costly to implement.

Agency reactions to a common geoscience data model for Canada were mixed. Most agencies thought that it would be good to have a common standard that they could map their data to if they wished to exchange data with another agency. The idea of adopting a common model for their local data sets was less enthusiastically received. There were many reasons for the reluctance, including the following.

- Models are difficult to implement, and time and resources are lacking.
- Knowing something about the data one already has (e.g., the meta-information) is a significant task in itself. That needs to be handled before even considering a common model.
- No common geoscience model has been tested for all the geoscience subjects of interest. It will take time to develop a common model, and it could involve costly mistakes.

- Once a common model is in place, it can be changed only through paperwork and committees. Immediate local needs would have to be placed ahead of standardization needs.
- Regions have special needs, so a common model would need to be very flexible or very complex to permit or support all regional variations.

In spite of an initial reluctance, most agencies did acknowledge that they would move toward adopting a common model if they could see that it handled the data in a manner that suited their needs. All agencies were ultimately practical.

5 Review of the Business Drivers

The primary purpose of a standardized data model is so multiple users can gain access to and understand the information gathered and maintained by various stakeholders in the CGKN. This section provides a high-level analysis of the business drivers for the CGKN.

- Need to have intra-sector and cross-sectoral connectivity
- Need to improve information management, access, and distribution
- Need to have information synthesis and analysis
- Need to support secondary and value-added usage
- Need to support non-traditional information applications

5.1 Intra-Sector and Cross-Sectoral Connectivity

The CGKN Portal must be able to support access to and from other earth science web portals. A good example of such a web portal is CORDLink. CORDLink is an Internet-based prototype digital library access mechanism (Figure 5-1) that is intended to integrate digital maps, images, and text on Canadian Cordillera geology.

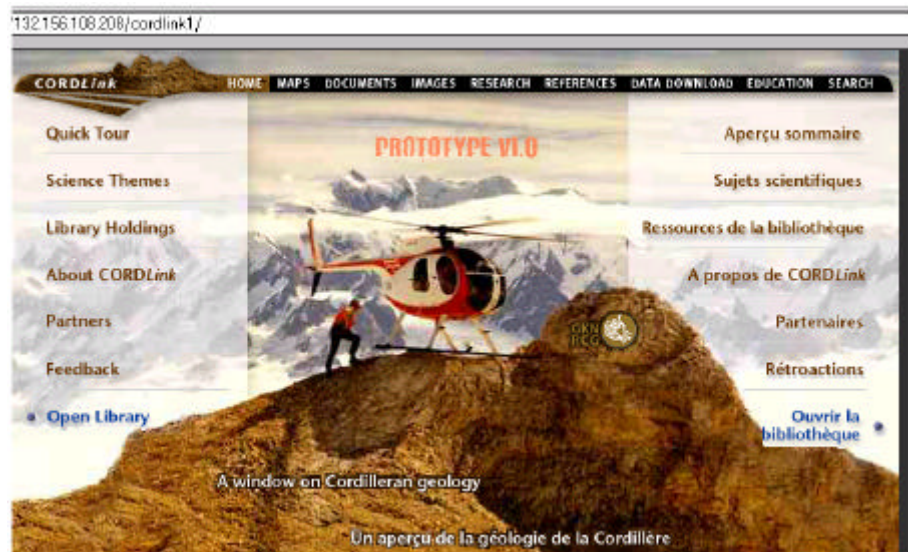


Figure 5-1: The splash page for the CORDLink prototype digital library portal.

As soon as possible, CGKN needs a single main portal for access to a data and services registry. The registry would describe what data and services are available through the CGKN portal. Users interested in geoscientific information would view an index map and subject categories through the main portal. When the user chooses an item of interest, the portal links the user to the actual content, which is held by its custodial agency.

In the fully developed CGKN, there would be multi-portal entry into the virtual webhouse. The user still can query a CGKN location map and subject categories, but the system also supports navigation to cross-sectoral data repositories. Likewise, other sectoral initiatives (e.g., CORDLink, GeoMatter, and NRIII) could navigate to data registered with the CGKN portal. The infrastructure will use standards-based interfaces and protocols along with cascading server technology developed through OGC for both the spatial and catalogue servers. To ensure that the CGKN portal is accessible to and from multiple Internet portals, the appropriate metacontent will be captured and propagated to CEONet and other catalogue servers.

5.2 Inventory of Available Data

Before data can be used, its prospective users must know that it exists. This implies that geoscience data that can be made available to those outside its immediate owners must be inventoried and properly catalogued. The best way to catalogue what exists is by using a standards-based cataloguing method that includes metacontent. The agency survey that was part of this study revealed that many agencies felt that collecting basic metadata for their datasets would be a significant but extremely worthwhile challenge that they were willing to undertake. Section 6 discusses metacontent as it relates to the CGKN requirements.

5.3 Manage and Distribute Data

The CGKN must provide a rational, corporate view of geoscientific information assets of all types. The status quo is a mixture of GIS, image, database, and file management systems that do not encourage integration or access. These heterogeneous environments are difficult to integrate and have high costs associated with training, support, maintenance, and upgrade. Information management requires the following.

- An extensible, scaleable data management platform capable of storing and managing traditional (tabular and GIS) and non-traditional (airborne geophysics, non-terrestrial satellite imagery, borehole logs) data.
- A transparent Web access using a dynamic catalogue of registered data holdings and an integrative geoscience data model.
- A decomposition of geoscientific data into data holdings that users can access, analyze, and report on directly and as needed, rather than resorting to pre-packaged data products (maps, reports etc).
- Data mining methods that use standard meta-information standards according to international standards developed by ISO TC211 and SC32.

5.4 Synthesize and Analyze Information

CGKN web portal must support the synthesis and analysis of information for the following geoscience purposes.

- Intra-sectoral studies and research that supports federal policy and directives.
- Wide-area, cross-provincial studies to develop an understanding of how geoscience practices affect global processes,
- Cross-sectoral data analysis to develop an understanding of economic impacts and benefits.
- Support to industry for exploration, planning, environmental assessment, and exploitation of resources.
- Supporting the planning for ongoing stewardship of Canadian resources for long-term benefit to our economic and cultural well-being.

This level of information integration requires a harmonized and extensible information model, as well as an open technical architecture that supports query and retrieval of information from many sources. All of these goals cannot be met immediately, but the tools and interfaces can be designed for extensibility so that the system functions can evolve along with user requirements. The initial CGKN Webhouse should include the following.

- A geoscience data model based on an enhanced version of the GSC/USGS NADM Draft v5.2 geoscience model. This model will continue to mature as development progresses. Established physical models such as the PPDM should be used to extend the applicability of the logical NADM model.
- A meta-information registry with a maintenance interface so data custodians can create, edit, and delete references to their data sources.
- An open, extensible web interface supporting query of spatial and attribute data, and meta-content.

- An open spatial database that can be used to store reference material for displaying the distribution of information assets.

5.5 Support for Secondary or Value-Added Usage

Although government geoscience data is captured in many cases originally to meet the information needs of a specific client, a wider audience exists and should be provided access. The CGKN Webhouse also needs to support the following types of commercial users and usages.

- Value-added resellers
- Information brokers and agents (print on demand, virtual commercial networks)
- Non-government organizations, such as environmental groups
- Geo-technical companies

These users fall into two categories. The first two user groups will be brokers who may add value to the information and then market the digital information to a wider user community. The third and fourth user groups may access the digital information from either a primary provider (the GSC, provincial or territorial agency) or a secondary provider. They would use the information in conjunction with other data for analysis related to issues such as resource extraction, environmental analysis, or sustainable development.

5.6 Support for Non-traditional Information Applications

Geoscience data is of great interest to non-geoscientists. Each of the following groups could benefit greatly from the ability to access geological and other earth science data on-line.

- Schools, i.e. students and professors
- Intelligent transportation systems
- Insurance and hazard evaluation

Canada's economy has great potential to expand in the new digital age. This can only be true if the educational system can be part of the "*Information Highway*". Currently, this is not the case because, even if they are connected, many schools do not have access to valuable information about Canada. The CGKN Webhouse could change that by providing geo-scientific information to Canadian schools.

A similar case can be made for hazard analysis in a country where natural hazards often relate directly to geology and geomorphology. Disaster relief and reduction of human suffering are intangible goals, but a system that can supply real information to evaluate slope stability, ground water flows, earthquake, or health hazards to reduce insurance costs is of measurable value. The CGKN Webhouse would make a concrete contribution to those goals.

5.7 Improving Overall Efficiency

Consolidation or integration can be a two-edged sword that reduces duplicated effort but in so doing, creates the overhead of committees and consensus. The goal is to capitalize on the first and minimize the second. We should expect to see increased efficiencies that stem from:

- increased commonality (look and feel) in systems design and utility,
- increased application of standards across data and systems,
- reduced duplication of effort,
- reduced duplication of data storage,
- reduced processing,
- a synchronization of update cycles,
- streamlining of operational procedures,
- an increased flexibility of product or dissemination choices, and
- less complexity in management.

Unfortunately, many geoscientific or geospatial warehousing initiatives focus on only the first two of the above elements. This has resulted in either massive repositories that overload even the largest servers, or the creation of silos of stale data. Many western countries have developed spatial data warehouses that revolve around a central portal or centralized authority that distributes information for a fee. These implementations have met with limited success for the following reasons.

- The systems have been unwieldy, forcing the data provider or prospective data user to expend more effort than previously required to acquire the desired information.
- The systems were limited in scalability. Once the achieved goal of widespread use of the system was achieved, the hardware could no longer support the user base.
- The systems were proprietary, forcing the user community to adopt a particular software solution. This adds costs and results in competition between agencies that are striving to provide information to widest possible set of clients.
- The systems were based on solutions where access was limited to tools that required licensing fees for each desktop interface that provided access to data. This may be acceptable to private industry, where the costs for services are easily recovered, but it does not work in the public sector, where much of the core geoscience data resides.

The landscape has changed with the arrival of the web-based delivery tools and powerful metacontent engines to allow distributed catalog services. Newer tools can be tuned to the needs of particular user communities, and also allow access to information sources to a larger audience.

Increasingly, multiple Web-enabled spatial data warehouses built on disparate technology platforms are able to be linked, even when they are a hodgepodge of formerly incompatible technologies. This phenomena is the beginning of a trend toward systems that logically integrate multiple physical databases into a larger logical whole, letting locally controlled systems be effective for a more global purpose.

The CGKN can take advantage of newer web-based delivery tools to design and deliver a system that avoids past pitfalls. The goal is to provide a rational, corporate view of geoscientific information assets of all types. One key to achieving this goal is metacontent

6 Discussion of Metacontent

Metacontent is often referred to simply as metadata. The term "metadata" can mean different things in different contexts. The information management community reacted to the ambiguity of the term "metadata" by introducing the additional terms metacontent and meta-information, and refining the use of the three terms.^{8,9} Metacontent is the generic term; metadata and meta-information are the two categories of metacontent that are related as shown in Figure 6-1.

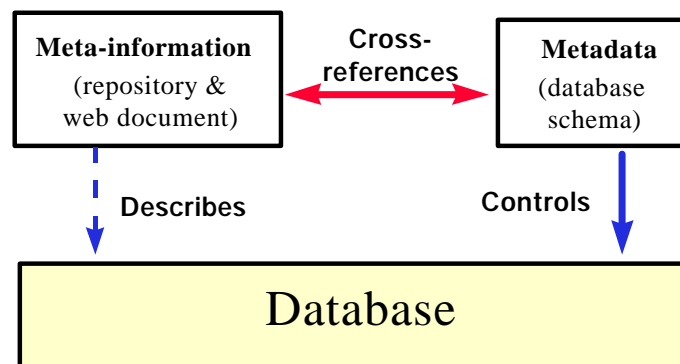


Figure 6-1: the relationship between data, metadata and meta-information.

Strictly speaking, metadata is the syntax contained in the data dictionary that defines the physical database schema. It describes where and how data are stored (e.g., replication, disk striping, and data partitioning), how to access data (e.g., indexing methods and topology), and when and how data are updated (e.g., version control, timestamping, and concurrency). Usually, metadata is defined by system analysts and data administrators. In relational database systems, metadata uses the Data Definition Language (DDL) component of the SQL data base language and is stored in the data catalog.

The distinction between metadata and meta-information is based not only on authorship and storage, but also on the medium of expression. Metadata involves technology-oriented syntactic descriptions; meta-information captures the semantics and pragmatics of human communication.¹⁰ Meta-information (description and/or

⁸ Podehl, M. (1993). User Requirements for Statistical Meta-information. *Statistical Journal of the United Nations Economic Commission for Europe*, 10, 2, 113-120.

⁹ Sundgren, B. (1993). Statistical Meta-information Systems - Pragmatics, Semantics, and Syntax. *Statistical Journal of the United Nations Economic Commission for Europe*. Vol. 10, No. 2, pp 121-142.

¹⁰ Kucera, H.A., and Flaherty, M., (1997), *A Spatiotemporal Backbone For Information Systems*, Spatial,

documentation) is essential because it provides analysts with details about the origin of the data and how they have been collected. In general *Metadata* permits transparent and error-free access to data, while *Meta-information* allows people to find data that meets a particular need, and judge its fitness for use. Figure 6-2 gives an example of how the BC Ministry of Environment has implemented their Data Registry based on these concepts.

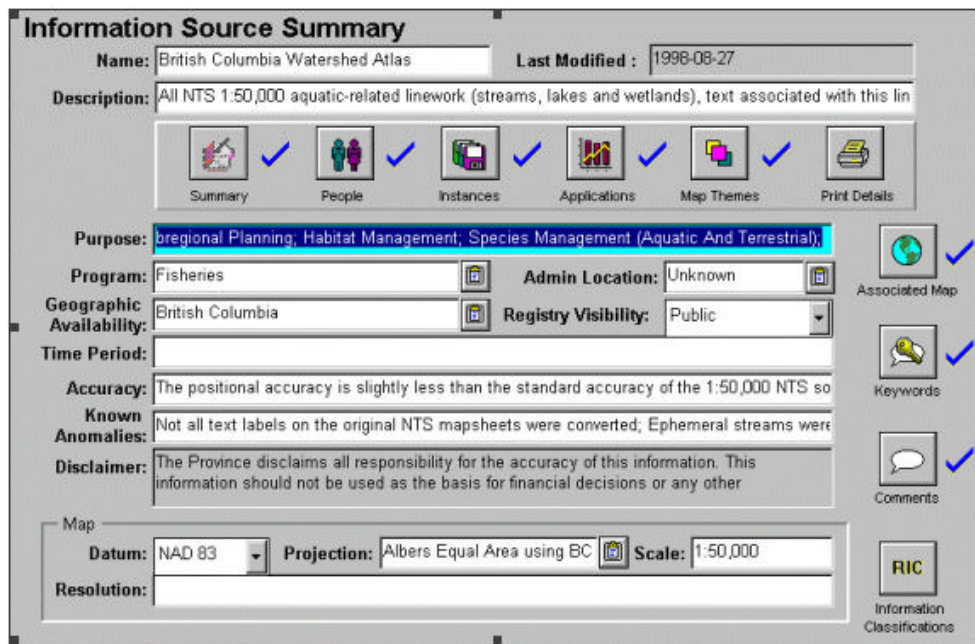


Figure 6-2: An example of browsing a data registry that supports common metacontent.¹¹

The concept of meta-information is very important to understand. People use meta-information every time they frame a question. An example of meta-information use is in the investigation of a community affected by potential contaminant drainage. To determine any causality an analyst must determine the spatial correspondence between the contaminant source and the water source for the populated places. S/he also must determine the groundwater characteristics of the underlying geology. To begin the analysis, s/he places parameters on age of information, source of information, accuracy, and completeness. Is the information s/he needs in paper (maps, photos or reports) or digital form (LANDSAT, well logs, or seismic surveys)? Does s/he need to be concerned with EPA records? Is there a link to legislation? Who owns or controls the land? What specifications governed data collection, analysis, and storage? These questions are answered by meta-information, not metadata. Is this a

Semantic, and Temporal Data Integration for Application in Remote Sensing and Geographic Information Systems, Monograph 47, Cartographica, Vol. 33, No. 1, pp. 95 - 105, Spring 1997.

¹¹ From Kucera, H.A., and Faulkner, A., (1998), MetaData + MetaInformation = MetaContent: Uniting Theory and Practice Using Oracle Designer, Oracle Open World Conference, Moscone Center, San Francisco, California, November 8 -12, 1998.

likely scenario? The recent non-fiction bestseller “A Civil Action”¹² chronicles a personal injury case against major US manufacturing companies and is based on exactly this type of investigation.

The importance of metacontent grows as the heterogeneity of data types and sources increase. To ensure that data are re-usable the agency responsible must describe their provenance and storage method in sufficient detail that users outside the source organization are able to assess the data's fitness for an intended purpose, and achieve access to the data without undue difficulty.

Figure 6-3 illustrates the roles of metacontent when a user interacts with a data store. A metacontent repository can have many functions.

- It assists database administration via version control and data sharing, it helps to control editing tasks that involve retrieving and changing elements.
- It stores engineering data and business rules to support application development and database design for system architecture.
- It provides meta-information for browsing and integration during decision support.

Not all metacontent repositories perform all these roles; a repository designer may consciously choose a subset of roles to perform.

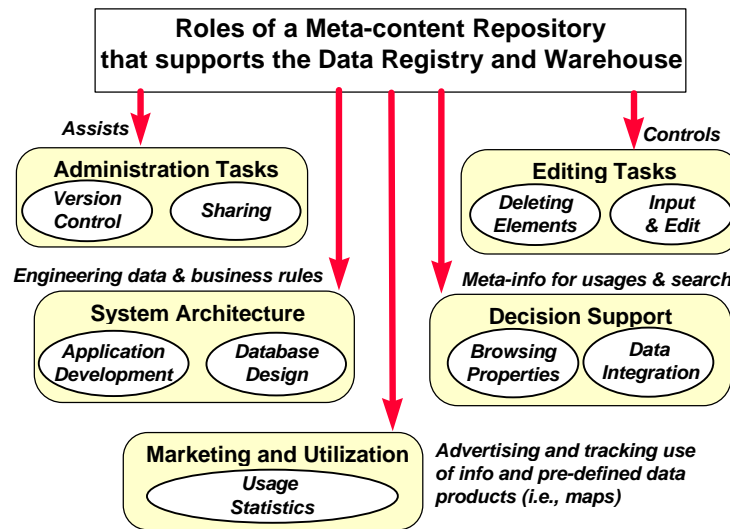


Figure 6-3: Roles of a metacontent repository (MELP, 1997).

The agency survey identified a metacontent repository as the most important first step for developing an integrated information system. The metacontent repository would deliver simple easy access (i.e., decision support) and a means to share system

¹² Harr, Jonathan: 1996, A Civil Action, Vintage Books

architecture Information to help design the next stage of common models and application interfaces. Many agencies felt that even collecting basic metadata for their datasets would be a significant challenge, but well worth the effort. This first step would require only simplified metadata or meta-information to make it a more manageable task.

7 Recommendations

"We need to take seriously the scientific issues of hooking data marts together, rather than arguing that these independent data marts shouldn't exist." (Ralph Kimball, 1999¹³)

This study examined the information modeling requirements associated with building, deploying, and distributing information through a national geoscience network. The survey and interviews with the participating agencies provided a good picture of the business requirements for storing and accessing data from a distributed knowledge network for geoscientific information.

Essentially all agencies agreed that a common Canadian geoscience portal to individual agency web sites would be the right first step toward an operational network. A user could enter the portal, query on location and/or subject, then be directed to the appropriate agency's or agencies' web site(s).

The portal concept also addresses some of the issues raised by agencies during the survey and interviewing process. Most of the territorial and provincial agencies stated that they must maintain an obvious local presence. A common portal that leads to an agency's local site enables an agency to participate in a national initiative while still maintaining a local profile.

At the workshop meetings held in Calgary on June 4 and 5, key requirements of the portal were articulated.

- Low cost of operation.
- Easy access to a registry that describes agency data holdings and provides a path to get to them.
- No training required.
- No installation on desktop.
- Open access to meta-information.
- Warehouse modeling & re-use capability, so each agency can participate in the growth of a common geoscientific model.

A picture emerged of a distributed linked network of data servers and services that communicate through a services registry. The services registry must have a common

¹³ Kimball, R., (1999), "Stirring Things up." *Intelligent Enterprise*, June 22, Vol. 2, No. 9, Miller Freeman.

metacontent schema that is agreed to by the participants in the network, while each linked database system can retain its unique information model.

Integrated data access is more challenging than an integrated data registry. Internet-based, enterprise-level spatial and aspatial data management is still in its infancy, so it is not straightforward to implement a system that can meet the requirements of a distributed network without some analysis and planning. An Internet-based system will almost certainly require the definition and use of:

- a shared services registry,
- a spatial/subject web client,
- a common logical model,
- common protocols (TCP/IP, HTTP, GLTP,) languages (GML, XML, HTML) and data transfer mechanisms,
- mechanisms for managing, querying and delivering the data, and
- open productivity tools and links to search engines to ensure access to all levels of the public and private sectors.

If these components are included in a given architecture, the result is a robust implementation that has dependable and quick access, regardless of the size and complexity of the database.

This section recommends a plan for developing the desired system. Given the disparity of capability and resources available across the CGKN, and the immediate need to see progress, we recommend that re-usable components be delivered in a logical sequence. The recommended plan has four sets of components that leverage existing applications, databases, metacontent repositories, and web servers in an incremental approach.

The architecture for implementing a web-based system that is upgradeable, and has re-usable components, must be based on a 3-tier technology architecture and a distributed-services model that uses open standards-based interfaces and protocols. Several implementation options are available. Figure 7-1 shows an implementation framework that is based on the OGC Web Mapping protocols, the ISO TC211 Distributed Services architecture, and non-proprietary web application-server technology. An architecture similar to this could be delivered using the component-based approach described immediately below.

Many components in figure 7-1 are based on international standards. The database could be one of many SQL based databases that support spatial extenders based on the ISO SQL/MM Spatial Specification. This list includes Oracle 8i, IBM DB2, Informix Enterprise server and Sybase. The extenders are commonly called “blades”, or “cartridges” and many use the concept of Well-Known Binary Objects as specified by OGC. Other standards-based components include the JAVA Application Program Interface, the data catalogs and the XML interfaces for reporting on any structured document, catalog or diagram.

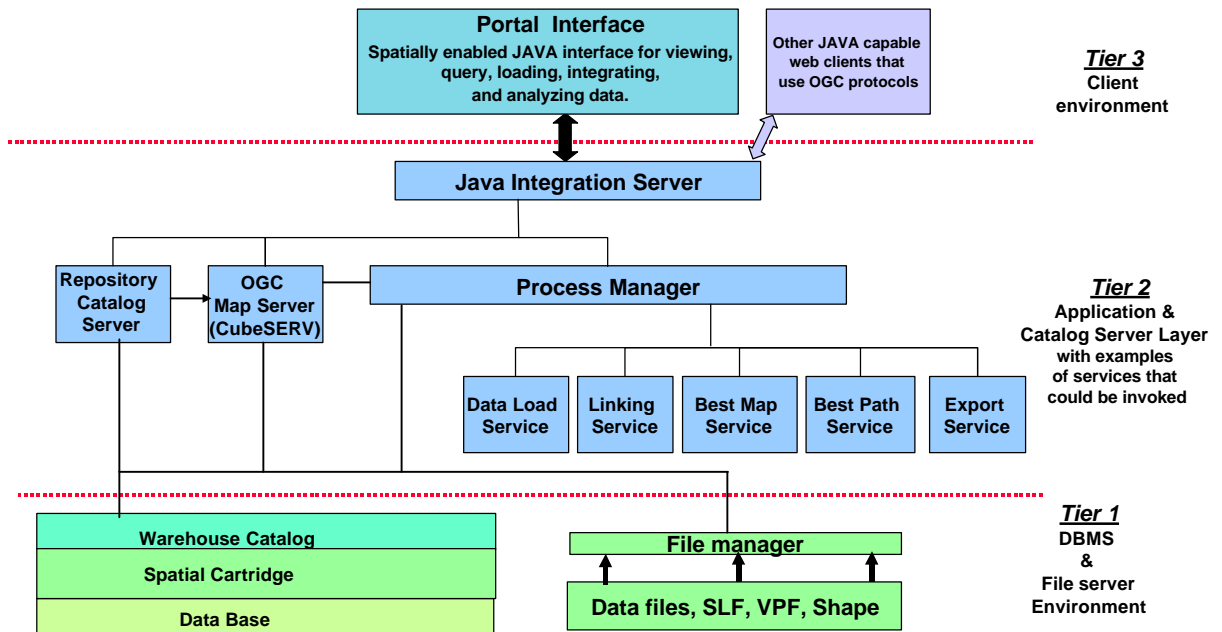


Figure 7-1: An example of a 3-tier distributed webhouse architecture.

7.1 A Multi-Component Approach for Development

This section provides a high-level summary of the recommended components for developing a Webhouse for distributing geoscience data across Canada. The components are also described in more detail in later subsections.

Component 1 is a first-generation operational CGKN portal that would deliver the following.

- Meta-information model (as described in Section 6).
- Single CGKN portal with peer-level links to participating agencies.
- A common map window linked to the meta-information model to spatially-enable CGKN.
- Shared services registry that links to the participating agencies' web sites.
- Small spatial database of shared common geographies that can be referenced by meta-information so users can easily navigate to the information assets in the CGKN.

Each agency could choose to make only parts of its site visible through the portal as services, or could have the portal redirect interested users to its site.

Component 2 would enhance the metacontent model for the services registry and would upgrade the CGKN Portal to take advantage of any portions of the common geoscience model that are available at this stage, and upgrade to use any useful new technology developments in spatial web delivery.

Component 2's CGKN would support efficient on-line navigation into time-variant data held by individual agencies without impacting their operational systems. This requires a more complete meta-information model for linking to services and would use the evolving common geoscience model (being developed concurrently in Component 3) to browse data assets held by partners in the network. The extensions to the core models would be the responsibility of the CGDMWG and fall within the tasks identified at the Calgary meeting.

Component 3 would develop a common logical geoscience model through interaction among agencies in the CGKN, industry, and other potential clients. It is recommended that the CGKN collaborate with standards bodies such as PPDM, ISO, and ODP to take advantage of existing and evolving efforts that may apply.

Component 4 would implement the complete common model for those agencies that have resources and a business case that supports the use of a common physical geoscience data model.

The four components are not mutually exclusive; concurrent work must take place on Components 1, 2, and 3. Components 1 and 2 can be completed relatively quickly by deploying existing technology within the stakeholder community. The need to show progress as soon as possible is met by making the CGKN web portal operational while work continues on refining a common logical model. Based on the experience of PPDM and the POSC Epicentre developers, a final Canadian geoscience data model will involve a longer, more incremental process.

Table 7-1 contrasts the four components. Table 7-2 shows approximate timelines for developing the components. This approximation is based on having adequate resources for optimal delivery, and a concrete set of confirmed requirements for each component. Care should be taken to ensure timely CGKN stakeholder input to the design and deployment process, and good conflict resolution so work is not delayed by disputes among stakeholders over requirements.

Table 7-1: Summary of the four components of development.

Component 1	Component 2	Component 3	Component 4
<ul style="list-style-type: none"> • Introduce and populate minimal meta-content model • Single CGKN portal. • Spatially enabled search • Services registry that links to agency web sites. • Common geographies used to navigate to information assets. 	<ul style="list-style-type: none"> • Enhance and populate extended meta-content model • Browse externally held data assets through single portal. • First iteration of the common model for a limited set of subject areas (e.g., bedrock, geology, soils, etc.) • Integrate recent technical improvements. 	<ul style="list-style-type: none"> • Develop common logical geoscience model. • Collaborate with other interested parties. • Extend the common model to include a wider set of subject categories with feature class definitions. 	<ul style="list-style-type: none"> • Implement common model • Assist agencies that are willing to adopt model elements.

Table 7-2: Approximate timelines of development.

ID	Task Name	Start Date	End Date	Duration	2000		2001			
					Q3	Q4	Q1	Q2	Q3	Q4
1	Develop Project Plan	7/1/00	7/21/00	15d	15d					
2	Initial CGKN Portal	8/21/00	12/21/00	89d	89d					
3	Improved CGKN Portal	10/15/00	8/16/01	219d	219d					
4	1st Iteration Data Model	8/21/00	1/19/01	110d	110d					
5	Data Model Implementation & Application Port	1/22/01	10/22/01	196d	196d					

The next four sections provide details on each component.

7.2 Component 1: CGKN Portal and Meta-information Model

Component 1 would deliver an integrated information delivery framework under a single distributed CGKN portal. The framework would implement a meta-information model and link it to a common map window.

The solution for Component 1 must achieve the following.

- ✓ Define a common portal to individual agency web sites.
- ✓ Support input from multiple formats and be able to output specified products if already offered by agencies.
- ✓ Build the subject catalogs in the services registry that can be queried using SQL, HTML and XML.

- ✓ Include a spatial database with common geographies that can be used to display and distribute information held by the CGKN partner agencies.
- ✓ Provide a service registration utility so agencies can manage and update the connection to their data.
- ✓ Be compatible with the operational systems within the various jurisdictions.

A key element of this phase is the expressiveness of the metacontent model, whose requirements are discussed immediately below.

7.2.1 The Extensible Metacontent Model

Given the complex and often interpretive nature of geoscience information, the extensible subject classification used by the metacontent model must allow multiple classification schemes for a given subject. The geological time scale is an example of a complex hierarchy with multiple interpretations. Figure 7-2 compares the Cretaceous period of the GSC’s 1995 Geological Time Scale Chart vs. the 1989 Harland time scale.

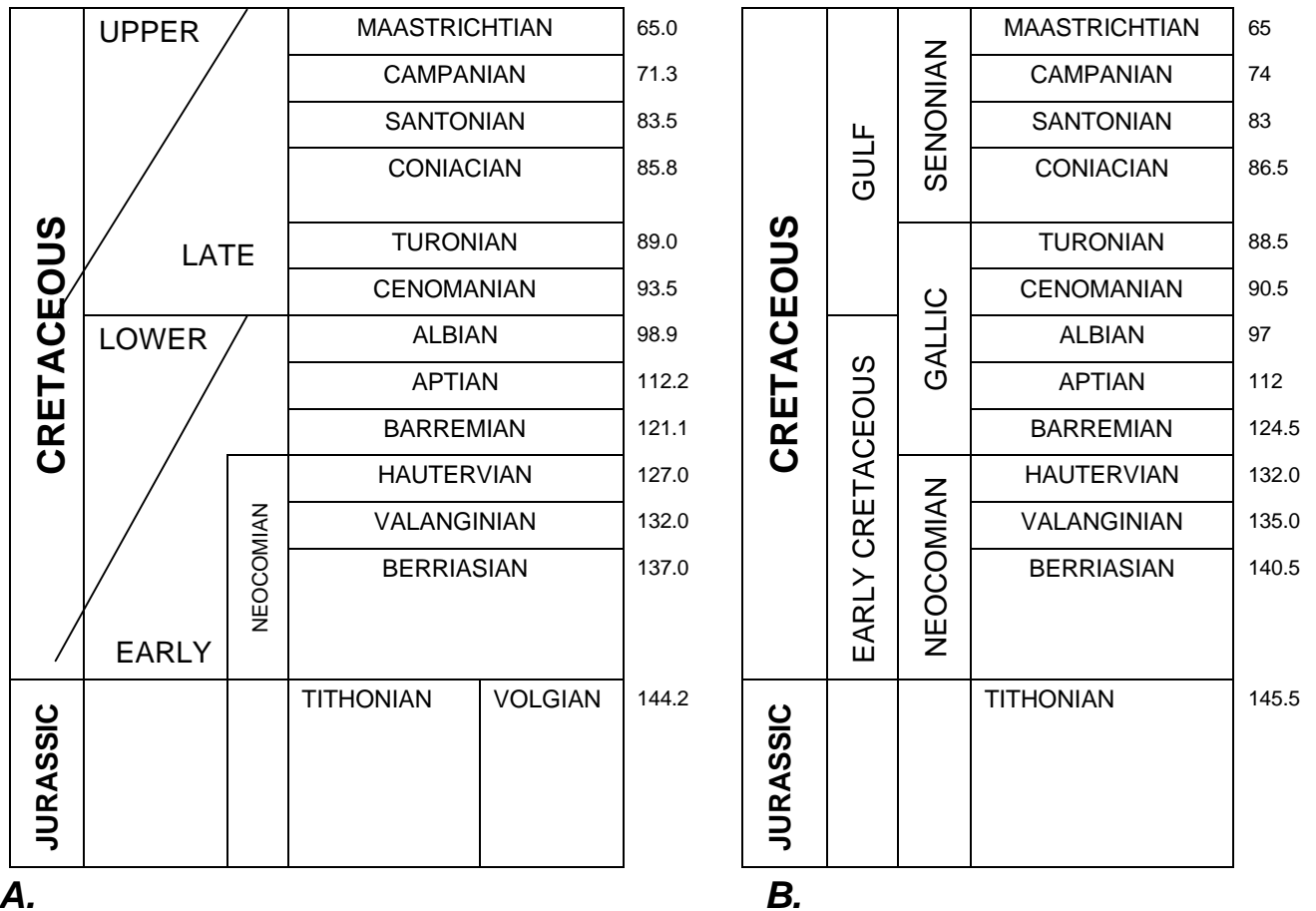


Figure 7-2. The Cretaceous Period as defined by (A) Okulitch, A.V. (General Coordinator).1995. Geological Time Chart. The National Earth Sciences Series Geological Atlas. Natural Resources Canada. GSC Open File 3040, and (B) Harland, W.B., Armstrong, R.L., Cox, A.V., Craig, L.E.,

Smith, A.G.& Smith, D.G. 1990 A Geologic Time Scale 1989. Cambridge University Press, Cambridge.

Although the stages of the Cretaceous Period are the same in both time scales, the ages (in Ma) and epochs differ. In the Harland time scale, the Senonian, Gallic, and Neocomian divisions are included but the first two are missing from the GSC time scale. In addition, for the top of the Jurassic, the GSC lists the Tithonian and Volgian as equivalents but only the Tithonian is included in the Harland scale. Even though these time scales are similar, there are sufficient differences that both scales should be supported. This type of variability occurs throughout the geosciences where interpretations are involved. As a result, it is critical that multiple interpretations be supported.

Among the four public data models reviewed, only the CORLink model and the ODP JANUS model provide a degree of support for multiple interpretation. The CORLink model supports interpretations by allowing the addition of tables. For example, to include multiple time scales, one can add additional *Descriptions* tables such as Stratigraphic Time Scale tables (Figure 7-3). The Stratigraphic Age table would need an additional attribute identifying which referenced time scale to use, and an additional table would be required to describe the overall time scale itself. This approach could be repeated for each subject. The difficulty of this approach is the propagation of tables. With each additional time scale, at least four more tables are added and when those are multiplied by the potential number of subjects, the number of tables to maintain is considerable.

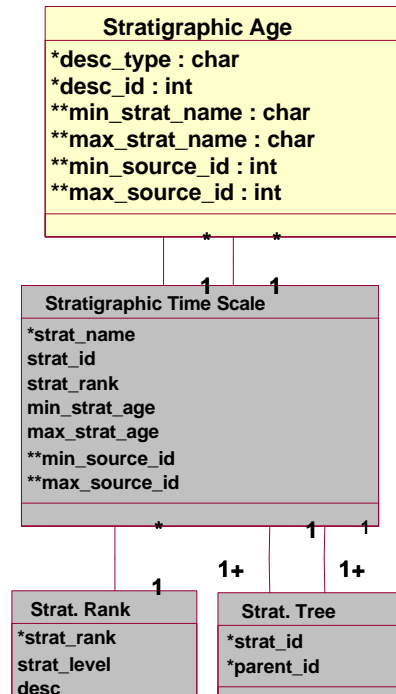


Figure 7-3. The Stratigraphic Age and related tables from CORDLink (NADM v5.2).

The ODP JANUS model deals with differing interpretations by using a concept approach. Any interpretive element (e.g. a geological age designation) is regarded as a concept which must be referenced to a publication or other designated source. In addition, any concept must include the name (or id) of the scientist adding the concept.

A possible approach to the extensible subject classification model for the services registry is to draw on elements from both the CORDLink model and the ODP JANUS model. A model that uses this approach is shown in Figure 7-4. In this model, the basic relationships among the tables in the CORDLink model are maintained, but they are generalized to the concept approach of the ODP JANUS model. This approach will keep the subject tables from multiplying when there are other interpretations. This could be considered a clarification of the abstract COA class structure in the NADM 5.2 model.

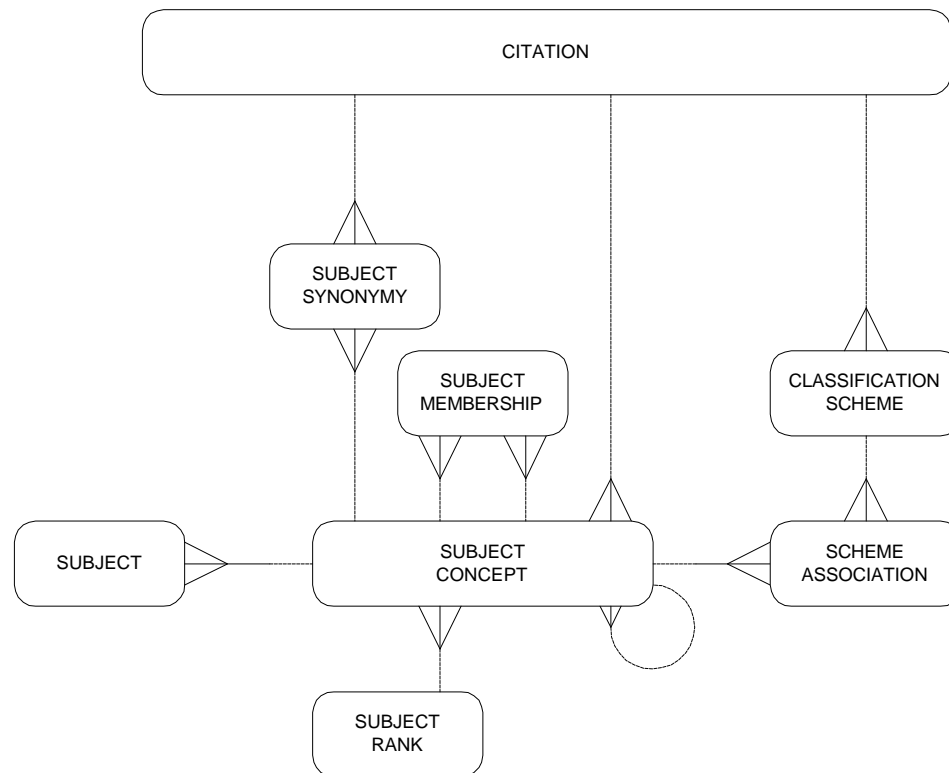


Figure 7-4. Extensible subject classification model.

The extensible subject classification model shown in Figure 7-4 is perhaps best explained by returning to a time scale example. If, for example, a paleontologist collects a fossil in the field and determines that that particular fossil is Maastrichtian in age, geological age would be the “Subject” and “Maastrichtian” would be a specific instance of the geological age subject. When that age is entered, the paleontologist can indicate which particular Maastrichtian s/he is using: the GSC version of Maastrichtian or the Harland version of Maastrichtian. The versions themselves are stored in the “Subject Concept” because an age is really an interpretation or “concept” sensu ODP. The “Subject Rank” for Maastrichtian would be “stage” and that stage is a part of Senonian Division in the Harland scale (which itself is a member of the Gulf Epoch, which is a member of the Cretaceous), or part of the Late Cretaceous in the GSC scale. The “Classification Scheme” is either Harland or GSC in this example, and the “Scheme Association” resolves the many-to-many relationship between concept and scheme.

The “Subject Synonymy” handles equivalencies. For example, the GSC time scale lists both the Tithonian and the Volgian for the Late Jurassic. To indicate that these are truly equivalent or synonymous, one or the other stage names could be placed in this table.

It is important to note that all of these elements are linked to a citation. Ideally, the citation is a published reference, or some other publicly accessible document that will serve as an authoritative support for the synonymy, scheme, and concept.

Initially, a simple subject hierarchy without the multiple interpretation could be implemented by omitting the schemes. However, it is important to include the schemes in the model design to ensure that multiple interpretations may be supported if desired.

This extensible subject classification model could be included in a larger, distributed internet-based service model, or meta-information model such as the one in Appendix I. The meta-information model shown in Appendix I is compliant with the evolving OpenGIS consortium standards and the TC211 distributed services model for geospatial data. It is important that any model developed for the CGKN Portal be as compliant with existing and emerging standards as possible.

The extensible subject classification model entities are indicated in Appendix I in green (or gray if printed). In looking at the model in Appendix I, a few additional entities are worth noting. The “Subject Class Usage” entity allows the user to know what is available for a particular subject. This entity links to the “Subject” and “Subject Concept” as well as “Data Set” and “Data Service” among others. The “Data Service” is particularly important as it identifies the mechanism that will be used to return data or information to the user. For example, if a user of the CGKN portal wanted to obtain information on fossils of the Maastrichtian age within Canada, the “Data Services” would identify which data sets are directly accessible (i.e., can be downloaded or queried within the CGKN portal) and which data sets require the user to be sent to a different portal (e.g., an agency’s local site).

7.2.2 Details on the Repository Browser and Services Registry

From a technical standpoint, the CGKN search, discovery and display architecture will require open interfaces to support thin clients and will need to be supported by some type of metacontent repository to support user interaction with the information.

To identify the design elements of the CGKN Portal, we have decomposed the requirements identified by the stakeholders into use cases (Table 7-3).

The key deliverables proposed in Component 1 are based on the use cases. They result in a design for an open repository browser that works with Netscape™ and Internet Explorer™, an extensible metacontent model, and the shared services registry.

Component 1 would result in a technical framework to allow designated content servers to register a service (data set or data transformation) with the common services registry. The web client can query the registry to discover what services are available for the subject of interest, and can invoke a selected service. Component 1 will provide the CGKN portal with a common map window linked to the meta-information model, and thus spatially enable the site. Figure 7-5 shows a mock-up of the CGKN Portal browser.

Table 7-3: Component 1's use cases for the CGKN Portal and Metacontent Registry.

Component 1 Use Case	Description
User launches browser	User logs in and browser opens to default view.

Component 1 Use Case	Description
User browses subject classes	The user interacts with a subject class hierarchy tree, and can browse subject classes by name, description, available data, and available services. The user can select subject classes for retrieval and display by the browser.
User browses spatial reference data	The user interacts with a map specification window to add and remove map layers from the browser. The browser renders the map view.
User selects backdrop layers	The user selects a spatial (map) backdrop for context. Features in the map backdrop cannot be queried, but the backdrop incurs little network overhead.
User pans and zooms the spatial view	The user adjusts the spatial view by panning and zooming.
User opens multiple browser windows	The user can open a new empty browser window, or can clone an existing open browser window.
User requests data service	The user selects a data service from a list of services available for a given data type.
User launches registration service	The user logs in and the system authenticates the user.
User registers new data service	The user registers a new data service by describing the data to be made available through the service.
User views all registered data services	The user views the data services that s/he has registered.
User edits registered data services	The user edits the information on data services that are registered by that user.
User deletes registered data services	The user deletes one or more data services that were registered by that user.
User tests URL for registered data service	The user tests the URL for a data service that s/he has registered.

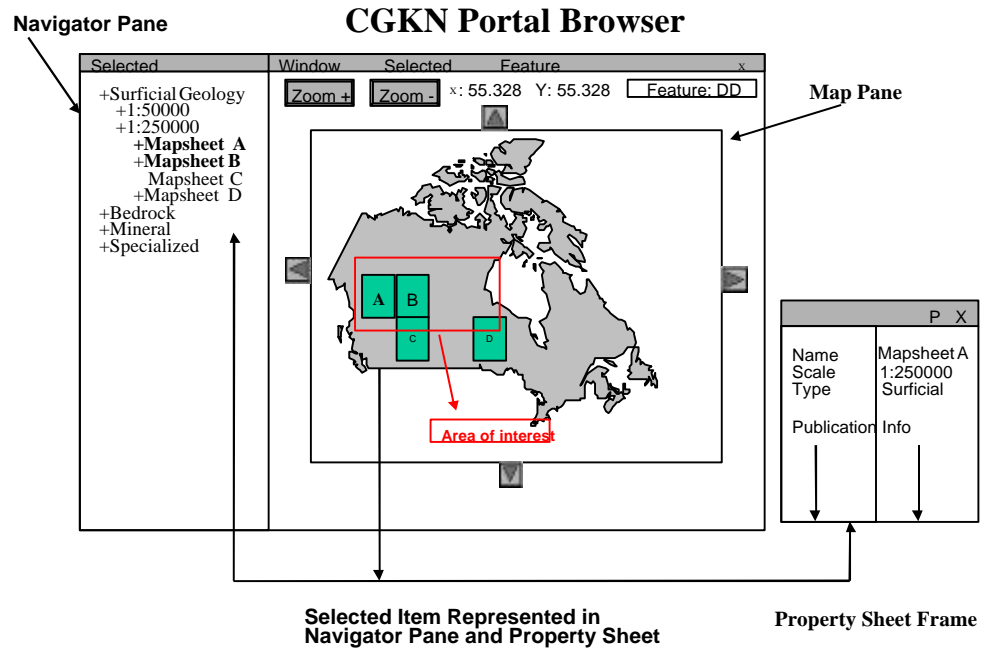


Figure 7-5: An example of the functionality of the Component 1 CGKN Portal.

The extensible subject classification model described above would be a critical component of the Services Registry of Component 1. The CGKN Metacontent and Services Registry would have the following components (Figure 7-6).

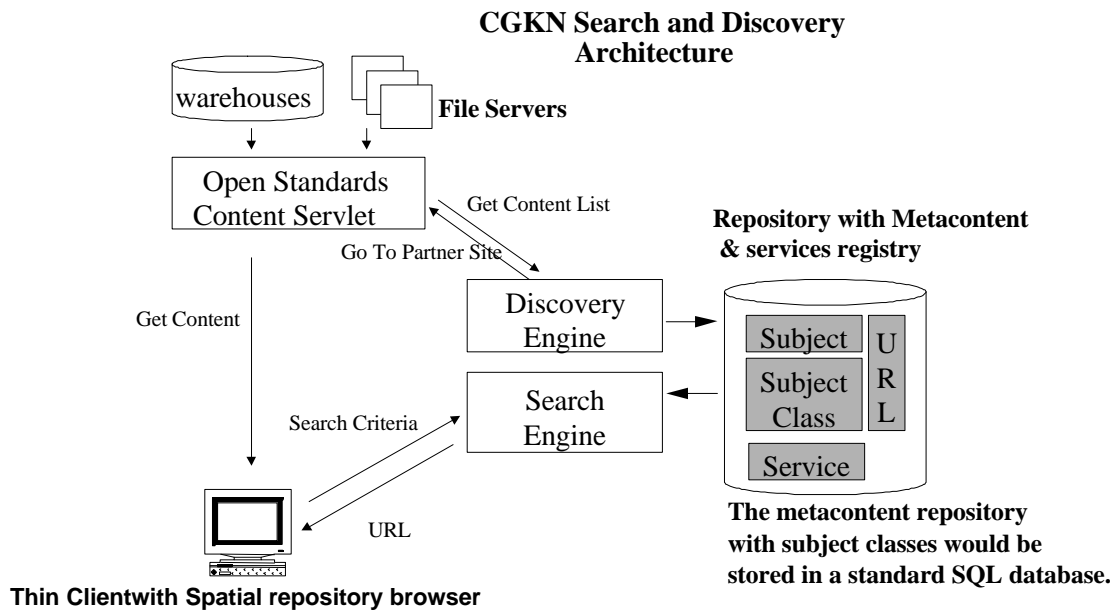


Figure 7-6: The proposed CGKN architecture for search and navigation to partner sites.

- An open standards-compliant Content Servlet (a program) installed at location on a web server that can provide two basic services. First, it can list the contents that it has knowledge of and that it has access to. Second, it delivers the content. For this to work, the Content List must include both the value and the type (subject and subject class), and it must be delivered in a structured format such as XML.
- A Discovery Engine that records both subject and subject class (value and type) into an index database. The Discovery engine provides a simple and uniform access to the repository content (meta-data and meta-information).
- Search Capability on a catalogue that supports queries on both value and/or type. Search results must be returned as URLs pointing to the content source page.
- A Browser Interface that supports the Search Engine Query Screens, the return links in an HTML page.
- A Thin Client Browser Environment that supports access to mime-compliant file formats on the desktop. Returned content does not need to be XML-compliant – it could be HTML, PDF, DOC, ASCII text, GIF, JPG, or any other file format.

Linking the meta-information to locations in the services registry will require the definition of a set of shared geographies to be adopted by participating content servers. Examples of shared geographies are provinces, townships, terrains, NTS map sheets and so on (Figure 7-7). Services and data holdings will be referenced to common geographies to allow a user to see what information resources are available in their area of interest.

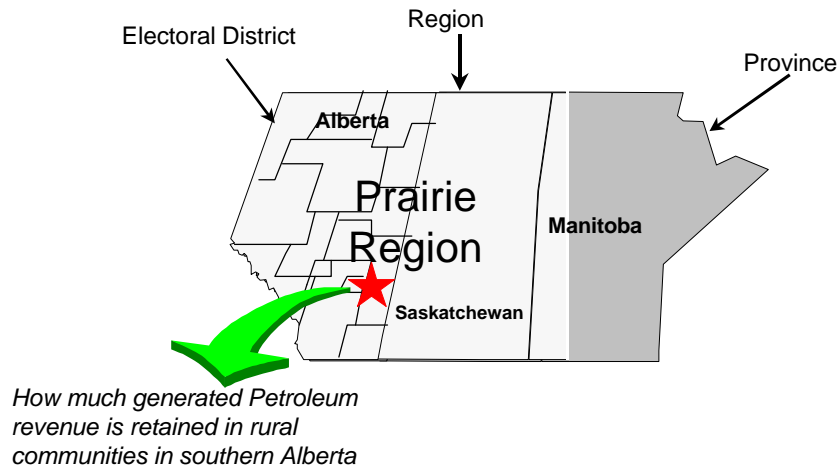


Figure 7-7: An example of a hierarchy of common geographies.

Figure 7-8 illustrates a conceptual architecture for the complete set of deliverables for Component 1. Note that once the information assets are discovered, the custodial agency may either make them available through the CGKN Portal or register a URL that directs the user to the host agency portal to retrieve the data.

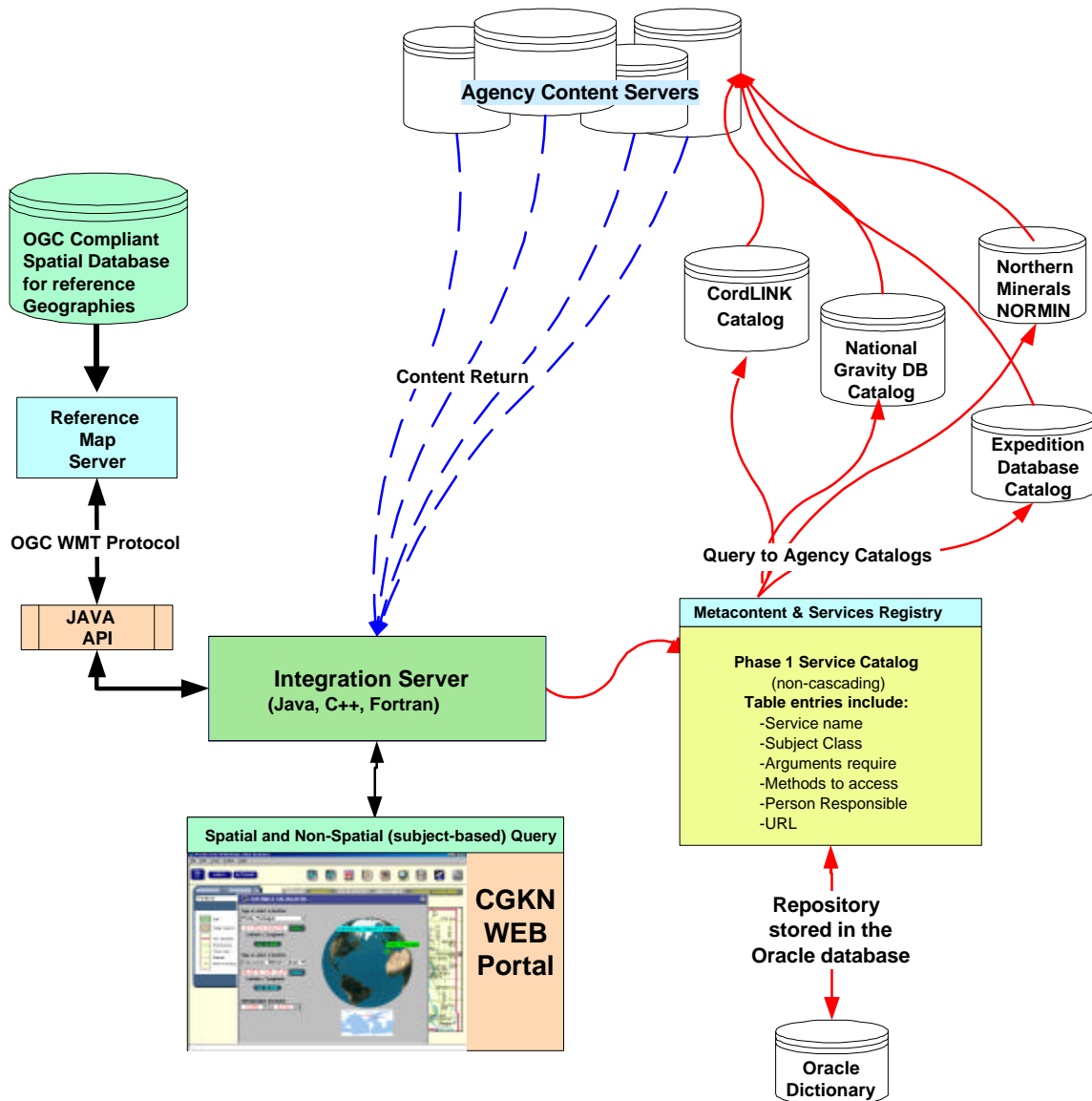


Figure 7-8: How the web interface would interact with the Metacontent & Services Registry to identify where information assets reside on the network, for the Component 1 CGKN Portal.

This architecture recognizes that information systems and data holdings will continue to be developed and maintained by individual agencies to fulfill their business mandates. The only spatial or geoscientific information that the CGKN portal must manage is reference geographies and subject categories that are used for presentation and navigation to distributed information resources.

Delivery time for Component 1 could be as little as 3 to 6 months, depending on the path chosen for implementation and the degree to which scope and requirements are established. Several of the technology pieces have already been purchased or

developed by member agencies in the CGDMWG. For development, Component 1 involves integration of the technology pieces, deploying the technology, technology transfer and training. Data conversion for Component 1 involves having each agency populate and maintain their portion of the Metacontent and Services Registry.

7.3 Component 2: Integrated Spatial/Subject Web Client

Component 2 will extend the basic CGKN Portal established by Component 1, to include direct on-line query of spatially referenced geoscientific data. To identify the key elements of the extended CGKN Portal, we again decomposed the user requirements into a set of use cases. Table 7-4 shows the Component 2 use cases that define extensions to the original functionality of the Component 1 CGKN Portal.

Table 7-4: Component 2's use cases for the extended CGKN Portal.

Component 2 Use Case	Description
User launches browser	User logs in and browser opens to default view.
User navigates to location using reference geography (e.g., mapsheets)	The user selects an area within Canada to query, and the browser displays that area with its mapsheet boundaries overlaid. The map names are listed in a separate window, sorted by subject and scale.
User chooses a product (this could be on a database or mapsheet(s) basis)	The user chooses an information product and views the name, scale, and publication information. The user can select the part of the product they wish to interrogate (i.e. a bedrock geology map for display of the vector map or the raster map image. The user also can choose a group of mapsheets).
User pans and zooms the map view	The user adjusts the map view by panning and zooming.
User filters vector data in the spatial browser	If an information products data is displayed by the browser, the user can remove or add feature classes from the display.
User views properties of spatial object	The user selects an object and the browser displays the names and values of properties on a "property sheet".
User pins property sheet	The user holds more than one property sheet open as he continues to view properties of additional objects.
User reviews the image of the map to see all properly represented items	The user views the raster image of a vector map to review its cartographic representation.
User downloads the product (e.g., mapsheet, geophysical log, image, magnetic survey etc)	The user chooses a "product" and a download format, and receives the data.
User places a "print on demand" order	The user submits a request to have a "product" printed.

In Component 2, a user accessing the portal would use a common map window to view geo-referenced geoscience information that is derived from individual agency data holdings maintained independently by the participating source agencies (Figure 7-9). This capability is achievable because the portal will have its own spatial database against which each content server may record data and implement services that are keyed by the feature identifiers of a shared geography. In this way, each agency may choose to allow access to their data through a registered service invoked against the shared geography even though the data is resident in their own content servers.

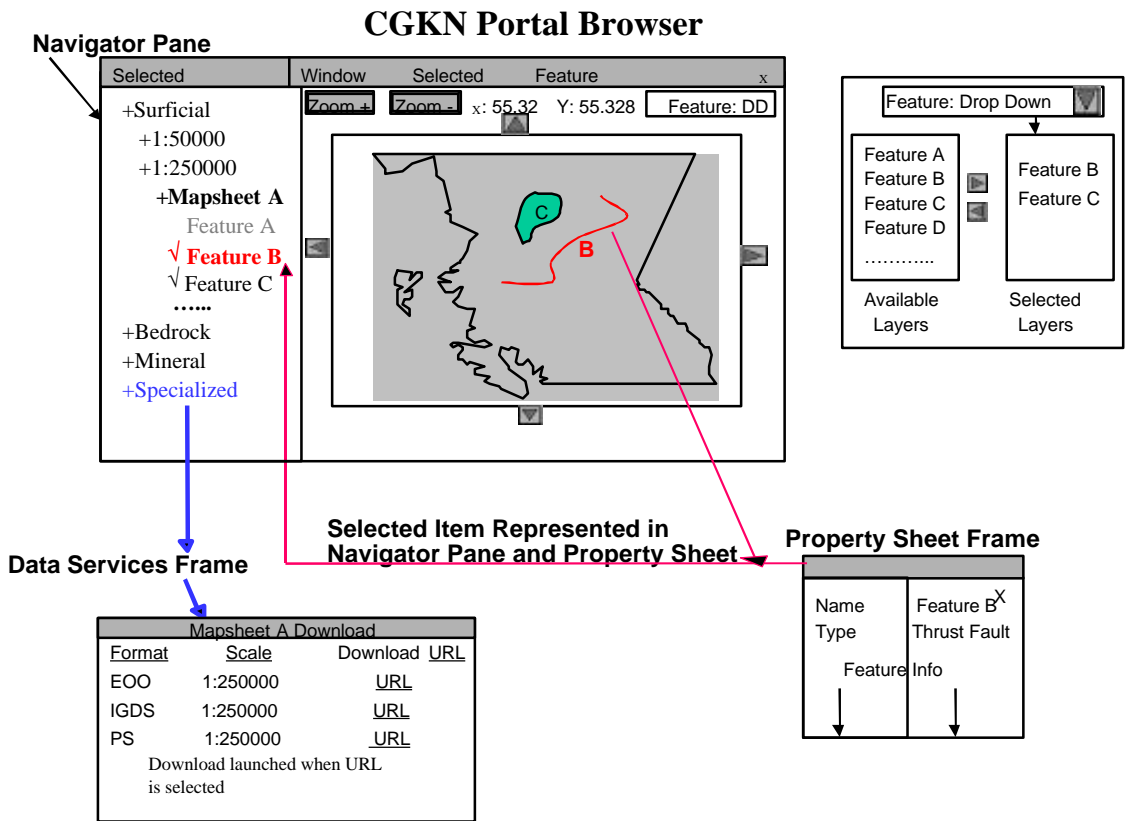


Figure 7-9: The Component 2 CGKN Portal interface would read individual agency data models to determine the location of subject information at the individual occurrence level. This requires a common logical model for distributed access.

Component 2 provides the following enhancements to Component 1. Figure 7-10 is a conceptual illustration of the connections and flows of the CGKN Portal following Component 2 enhancements.

- ✓ A capability for multidimensional query of non-resident data directly from the CGKN Portal. Users will have a choice of which rows and columns to display, they will be

able to filter on specific values of dimensions, and they will be able to select which measure to display.

- ✓ The ability to have “chained” services, where a user can choose an information resource and request a sequence of services to operate on the information. For example, the user could choose a data source, and have it translated to a particular data format, re-projected to a new coordinate system, generalized to the appropriate scale, then encrypted and transferred to the client machine.
- ✓ An interface to display data in the browser in a chosen geographic context with the ability to re-project and choose legend characteristics for presentation. The design created for this project would not approach a map production capability, but it would improve the graphic presentation of data being viewed in the browser. Extensions to the design would be assessed in subsequent phases. Ideally, it will be possible to integrate as services the common computational steps in the scientific data presentation and analysis life cycle: data retrieval, analysis (Kriging / interpolation), presentation (selection of colours and symbols for choropleth, contour, and other maps), and Internet publication (via gif, postscript, or screen print).

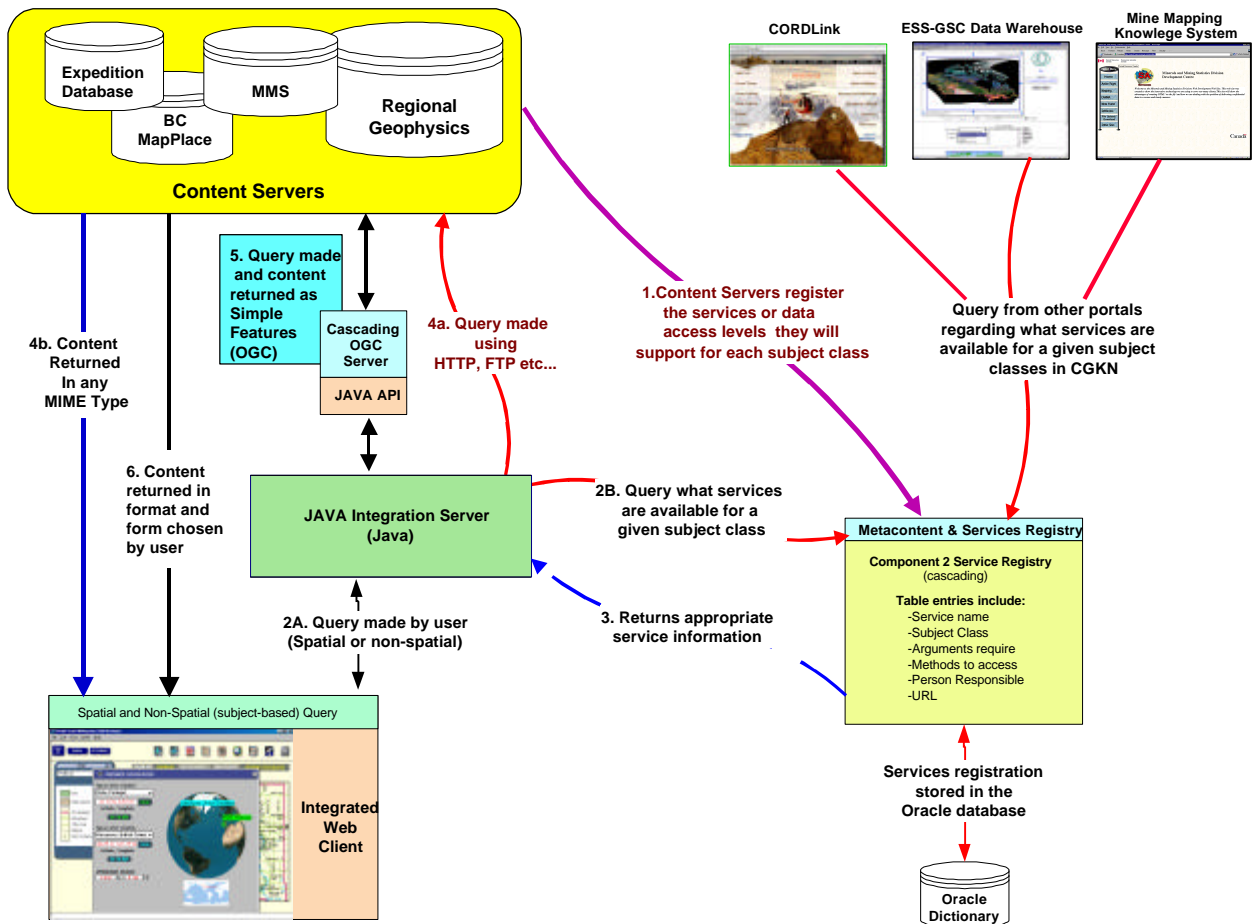


Figure 7-10: The Component 2 architecture displays subject and spatial data from the sources maintained by individual agencies.

The Component 2 architecture is based on the evolving OGC and ISO TC211 standards, and on the more mature ISO SQL Standard and W3C consortium web specifications. Standards compliance reduces risk and ensures that Standards-based Commercial Off-The-Shelf (SCOTS) tools are available to build and maintain the network (Figure 7-11). The technology is not based on untried research; it reflects ongoing system delivery projects undertaken by Holonics and by others for clients in the US and Canada. The architecture is also in line with the architecture being developed through the Information Interoperability Institute (III) on behalf of the CGDI Technology Advisory Panel.

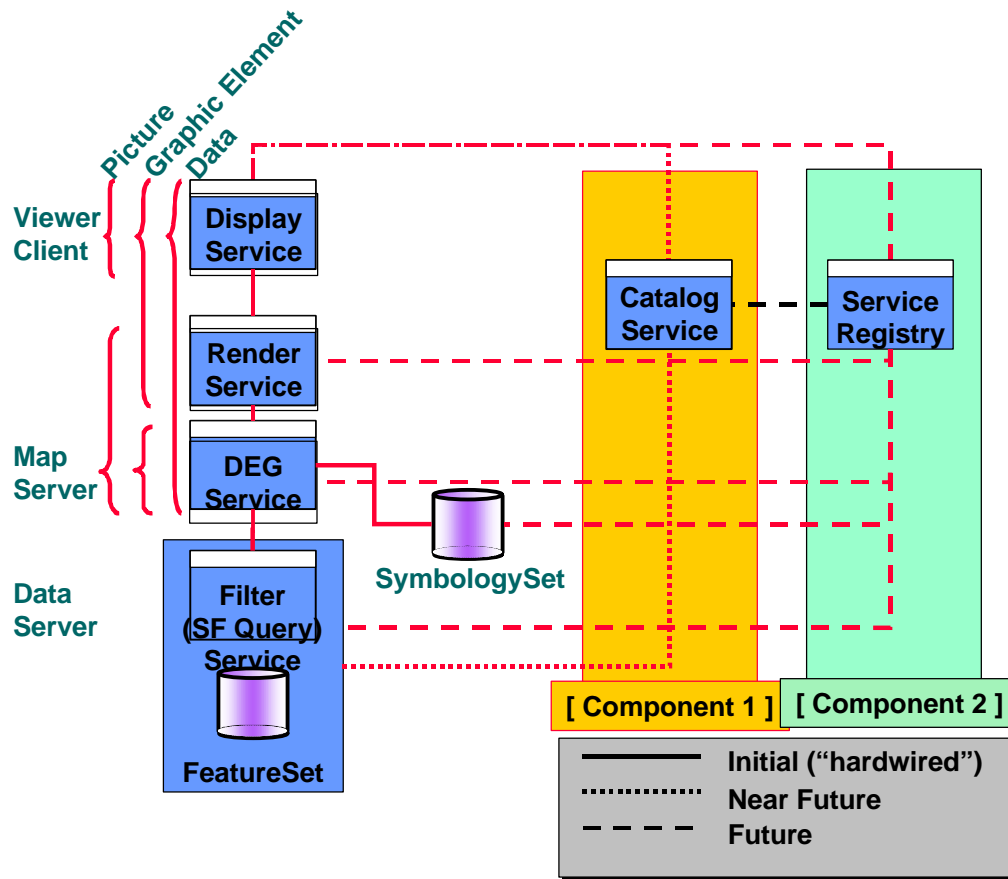


Figure 7-11: OGC web specifications and their relationship to the CGKN.

Several early adopters of technology with Natural Resources Canada (NRCAN) have been prototyping components of this architecture (ESS Data Warehouse, CORDLink, GeoMatter and others) and this experience can be applied to deployment of an operational network. Given appropriate resources for the stakeholders to prepare and register their information assets the elapsed delivery time for Component 2 should be less than 9 months.

7.4 Component 3: Develop a Common Data Model

The third component focuses on a common geoscience data model. Although this component can proceed in parallel with the other two, it will be important to give the first two components priority as they are the least controversial and most easily accomplished. A subset of a complete geoscience model should be developed to allow early deployment of the CGKN.

Section 2 stated that no single public data model currently fits the needs of the geoscience community in Canada. In addition, participating agencies are at many different levels of capability for IT development. Some agencies are just starting to

develop systems (e.g., Nunavut) while others have essentially corporate systems (e.g., Ontario & Quebec). Those just starting are likely to benefit most from a common model while those with large or corporate systems already have invested in their current systems and are less likely to make significant changes in the near future. As a result, it is impossible to propose a model that will suit all concerned.

Rather than try to develop a model for all areas of geoscience, or even the six broad categories defined by the CGKN, it is suggested that efforts would be better focused on areas for which established models are already in place. These include

- geological maps (NADM v4.3, CORDLink, GeoLegend, GSC Warehouse-GEMS),
- minerals (MINFILE, NORMIN.DB, MODS, etc.),
- geochemistry (MultiDivisional Data Model, PPDM, ODP etc), and
- hydrocarbons (PPDM).

It is recommended that technical teams be established for each geoscience category and that the teams include members from participating agencies and representatives from industry. The mandate of each team would be to produce a model that complements the models produced by other teams. The models should be drawn in a shared environment to ensure conformance.

Component 3 should achieve the following.

- ✓ Prioritize geoscience categories for data modelling effort.
- ✓ Identify models to be examined in detail within each category.
- ✓ Establish technical teams for each geoscience category. Technical teams should include both geoscience agency and industrial members. Ideally each team should also have a data modelling specialist to ensure the team adheres to fundamental data modelling concepts and designs. All data models should be hardware and software independent.
- ✓ Choose a common language and tool to use in the data modelling effort, for example entity-relationship (E/R) diagrams and Oracle DesignerTM or the Unified Modelling Language (UML) supported by Rational, Select, and others.
- ✓ Identify the attributes or entities common to all geoscience categories (e.g. location) and ensure that at least one of the technical teams has the mandate to develop and maintain that core area.
- ✓ Establish a reporting mechanism for technical teams on the CGKN portal so that all agencies and industry partners can be kept apprised of progress.
- ✓ Have a clear project plan outlining milestones and deliverables for each technical team. Funding should be in place to ensure that technical teams can meet the deadlines outlined in the project plan.

Completing the geoscientific abstract model will take time. But a first iteration based on the current NADM with some enhancement should be ready in 6 to 9 months to allow the integration of existing applications in Component 4.

7.5 Component 4: Implement the Common Data Model in Stages

Component 4 will implement the first iteration of the data model (developed by Component 3) under the CGKN portal and will include tasks such as porting existing applications like GeoLegend and GeoMatter or develop a new application for browsing and reporting for the web. While both GeoMatter and GeoLegend can be thought of as primarily data entry and editing tools, they represent an investment in the development of an interface that is meaningful to geoscientists rather than mappers. This investment is important if the audience you are trying to reach is the geoscientific community.

A detailed plan for Component 4 should be developed only after Component 2 is complete, since the technical requirements for Component 4 will be better known at that juncture. Part of the plan should include an evaluation of requirements against real near term and long term costs of distributed access tools. The means and costs for web query reporting and transactions (query, edit, update, delete) are very dynamic and there may be much better options available than the current crop of proprietary tools such as ArcViewTM or MapGuide^{TM14}.

8 Conclusion

This report presents a rationale and an approach for developing an integrated information sharing portal for the CGKN. The recommendations draw on formal and informal discussions with the CGKN and CGDMWG participating agencies, database developers, first-hand experience in system development, close monitoring of the debates surrounding international standards, and focused review of the database and information management literature. Based on this synthesis, it is recommended that a geospatial framework be built to exploit existing and developing technologies in a way that minimizes risk and shortens time for delivery of working systems.

If undertaken with firm commitment, the CGKN will become the integrating force behind the sharing of geoscientific information across Canada. The CGKN will provide the following.

- ✓ Extensibility of the underlying data model to multiple dimensions, scientific applications and geometries.
- ✓ Query flexibility to support user-defined searches over a continuous and seamless database of geoscience data.

¹⁴ ArcView and Mapguide are products of ESRI and AutoDesk respectively.

- ✓ Scalability to support additional data volume and changing data granularity.
- ✓ Open access through the use of standards-based web-enabled technology.

These components can provide an open N-Tier architecture that can be safely extended and modified as spatial application and database technology mature over the coming decade while meeting expressed needs of the participating agencies surveyed on behalf of the CGKN.

9 Appendices

Appendix I. The extensible subject classification model imbedded in an ISO/TC211-OGC meta-information services model

