Review Document

Evaluation of Candidate Models for GeoScientific Data

Draft: Version 0.3

Date: March 31th, 2000

Executive Summary

Canadian policy issues are increasingly being driven by national and international commitments in the areas of climate change, biodiversity, resource use, sustainable development and trade. Provision of information on status and management of Canada's natural resources is essential to put Canada in a favourable position to address global and national resource issues. Global change, biodiversity, and the sustainable development of Canada's economy are tightly linked to production from agriculture, forestry, minerals and metals and energy sectors. Creating government policy and programs to administer, foster and promote growth requires that staff have timely and efficient access to data resources. The Geoscientific community is evaluating several industrial and scientific information models in order to develop an operational common model for GSC, and NRCan as well as provincial and territorial geological agencies.

The evaluation of the candidate models is geared toward extracting and propagating the essential components of the existing model while at the same time making recommendations for improvements that will result in a more generally applicable and extensible new data model. The Goal is to increase the ability to integrate, manage and access information through distributed web-centric information access. The guiding principles of the design of the holistic data model are:

- The model must allow information to be located or captured easily,
- The model must be easily updated and managed, to allow tracking through time,
- The structure must be flexible so a source can be synthesized with other information sources to create a knowledge base for planning and implementation of operation programs,
- The model must be manageable, useful and complete with the ability to be improved and adapted over time so that the system can grow with the increasing complexity of data resources.
- The model must be easy to populate and update so that the data providers will keep it current and so that access over time continues to be transparent to the users.

To achieve this the model must be:

- based on open industry recognized standards (such as TC211, POSC and PPDM)
- extensible to allow growth descriptions of information assets and user groups
- capable of supporting application and services interoperability, allowing plug-and-play modules and services to be added.
- easily managed and populated (potentially in an automated way) by the information custodians (like information, metainformation should be captured once and managed at its source).

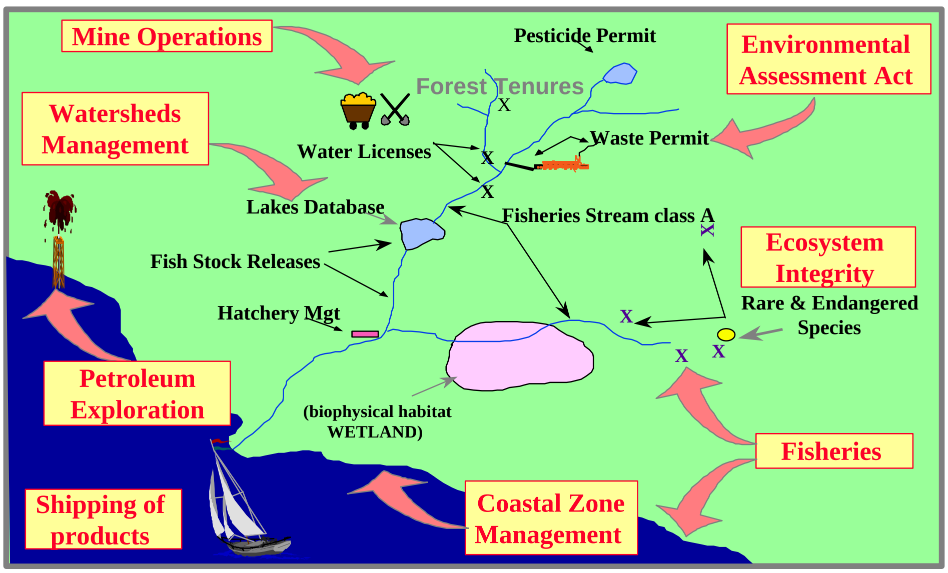
Four key candidate models were examined with these points in mind. These models are: PPDM, NADM v5.2, POSC Epicentre, and ODP JANUS. None of these models meet the requirements outlined above. Additional models will be evaluated during the survey phase of this project in consultation with the stakeholders.

Table of Contents

Data Model Review

1 Introduction and Purpose of Document

Policy issues are driven by international commitments in the areas of climate change, biodiversity protection, reporting on criteria and indicators of sustainable development and trade as well as national obligations for reporting to the public. Canada requires an information model that will support systems that can provide timely and accurate spatial and non-spatial natural resources information to the public and to policy-makers.



2 Overview of Objectives and Scope of Model Review

It is intended that a common geoscientific model will enable integration of data holdings and information delivery frameworks currently implemented in information systems within each sector. Merged information content will be derived from the underlying database systems managed within each sector. Access and integration of holdings will be managed through catalog services that will access information based on documentation stored in a common holistic model.

This draft document provides a first evaluation and makes tentative recommendations on the use of four key candidate models: PPDM, NADM, POSC Epicentre, and ODP JANUS. There are numerous other models worthy of evaluation. These include, but are not limited to, BASIN, MINFILE, and FieldLog among others. During the survey phase of this project, further information will be gathered on these and other models as put forward by the stakeholders.

In evaluating the four key candidate models, the main focus of each model will be noted. A model with a strong industry focus responds to quite different requirements than one with a strong scientific focus (Figures 1 and 2).

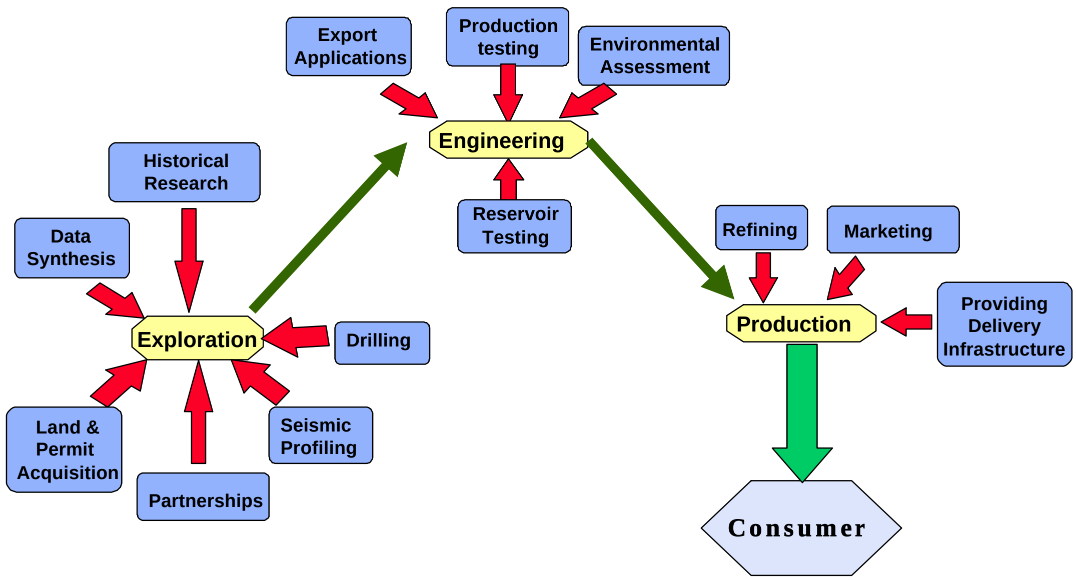


Figure 1. Industry focus

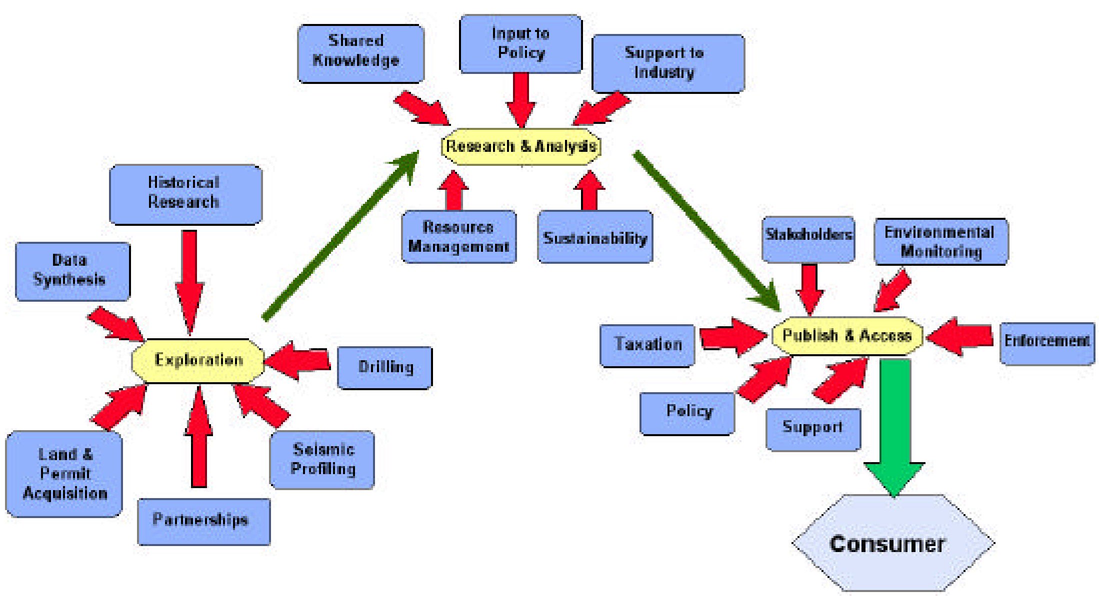


Figure 2. Science focus

A Canadian Geoscience Data Model must support both approaches. As a result it is essential that the model be developed through close cooperation with all stakeholders including key users groups, and providers of data.

3 Evaluation of Candidate Models

3.1 PPDM

The Public Petroleum Data Model Association is an independent, not-for-profit association representing over 100 oil and gas companies, vendors, and regulatory agencies worldwide. PPDM delivers 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.[1] The model is well established and maintained.

3.1.1 Analysis of the PPDM Candidate Model

3.1.1.1 Overview

The model is extremely well suited to its oil and gas industry audience but lacks the breadth essential to a core holistic model for the Geoscientific community. PPDM is dominated by well and wellbore information and associated supporting entities and thus supports its target audience very effectively. Geoscientific topics that do not specifically relate to oil and gas industry needs are not well specified. For example, only very limited paleontological and geological elements are included.

3.1.1.2 Detailed evaluation

- **Geological maps (GIS and attribute databases)**

PPDM records locational information about a node, point or vertex along with the map projection in the NODE table. Subtypes of the NODE table record locational information as expressed in the Carter Survey System, Congressional Survey System, Dominion Land Survey System or National Topographical Survey System, European Survey System, Offshore Survey System as well as several specialized American systems. The PPDM Spatial Workgroup has worked to integrate the spatial information in PPDM generically with Spatial GIS engines, such as ESRI SDE™ and SDC™ (now called Oracle 8i spatial). The model revisions are included in the beta version (PPDM Beta Model v3.5b).

Geological attributes are associated with well bores via the 'Lithologic Interval." Structures, porosity, diagenesis, rock color, grain size, and other rock type components

[1] PPDM-Public Data Model Association web site http://www.ppdm.org/ Modification date: March 10, 2000.

such as fossils are included. These entities are very focused to the industry audience and thus are not extensively specified.

- **Minerals/hydrocarbon databases**

  Minerals are poorly represented within PPDM but hydrocarbons are extensively modeled. Tables such as WELL_OIL_ANALYSIS and WELL_OIL_VISCOSITY record hydrocarbon data directly, while other tables such as WELL_PAYZONE, WELL_TEST, WELL_TEST_ANAL, and WELL_PROD_ZONE provide additional details on hydrocarbon related elements. Many of the "WELL_" suite of tables contain attributes associated with hydrocarbons.

- **Geochemistry/geophysics databases (raster and vector)**

  Geochemical analyses of oil, gas, and water form a portion of the PPDM model. Other types of geochemical data are not extensively handled.

  Approximately 20 entities fall within the seismic grouping of tables. Seismic energy sources (SEIS_ENRGY_SRCE), seismic lines (SEIS_LINE), amd seismic arrays or receivers (SEIS_RECVR_SETUP) are all included within the PPDM model. Extensive well log attributes are also included.

  Other types of geophysical data, such as paleomagnetic information and interpretive 3-D and 4-D models, are not currently part of PPDM.

- **Assessment reports and associated databases**

  The storage of a complete assessment report is not directly handled by the PPDM model. However, information concerning well licenses and rights plus production histories (daily and monthly) are included within the model.

- **Bore hole databases**

  PPDM has very complete coverage of petroleum borehole data. Most of the model deals with various aspects of the well.

3.1.2 Software independence and availability of support

- PPDM is well adapted to a relational platform and thus can be implemented in most any relational or object-relational environment.
- The PPDM model is available to members only. The cost of membership varies with the size of the organization (see http://www.ppdm.org/feestr.html).
- Numerous industry clients have applied the PPDM model successfully. Developed software remains proprietary.
- PPDM continues to expand the model to encompass other key exploration and production business processes. Special workgroups (e.g. the Spatial Workgroup, the

Stratigraphy Workgroup) have been established to extend the model in defined areas.

3.2 NADM

The 5.2 Draft of NADM was developed by the team leading the CORDlink initiative. CORDLink is an internet-based digital library access mechanism that integrates digital maps, images and text on Canadian Cordilleran geology. The CORDLink vision was to provide a comprehensive, flexible resource for scientific and non-scientific users of geoscience information. To make the prototype widely applicable for geoscience data, significant effort in collaboration with the US Geological Survey and the American Association of Sedimentary Geology has taken place. The resulting CORDLink data model serves as starting point for a wide variety of users of geoscientific information.

The CORDLink Data Model 6.0 was developed by Brodaric, Journeay and Talwar in response[2] to a review document produced by HydraSpace Solutions Inc.[3] In their response document, Brodaric et al note that a "scientific model" based approach is both a high-level requirement and an essential task. The CORDLink Data Model 6.0 is a good first step along this path. The authors note that this model is incomplete in terms of subtypes, roles, and class hierarchy development. Given the incomplete nature of CORDLink Data Model 6.0, this review will be confined to draft version 5.2 except as noted.

3.2.1 Analysis of the NADM Candidate Model

3.2.1.1 Overview

The candidate NADM v5.2 model is well structured and, although it is currently very limited in scope, it is designed to be extensible. The "scientific model" based approach is very well suited toward the display and dissemination of information as a cohesive set of products (e.g. maps) as well as the ability to represent interpretive earth science models through its Object Archives.

NADM v5.2 is effective for its intended audience in the area of geoscience modeling and geological maps. Entities or objects specifically directed toward disciplines that are largely sample based (e.g. paleontology and geochemistry) are not included in the model (or in the proposed version 6.0). Given the difference in requirements between the modeling approach and a sample analysis, it is likely that NADM v5.2 (and 6.0) will need to be significantly modified to accommodate these other disciplines.

3.2.1.2 Detailed evaluation

[2] Brodaric,B. M.Journeay, & S.Talwar. NRI3 Open Spatial Library Data Model Review. February 10, 2000.

[3] HydraSpace:"Developing an Open Spatial Library Model: Review of the CORDLink-GSC/USGC draft data model version 5.2." Draft: Version 1.0, January 4, 2000.

- **Geological maps (GIS and attribute databases)**

  The spatial aspects of the NADM v5.2 data model are very well defined. Objects such as the "Cartographic Object" and others within the Legend module make this model particularly well suited to geological map production.

  NADM v5.2 uses Singular Object Archives and Compound Object Archives to represent a variety of geoscience features. These object archives serve to associate a spatial feature with the descriptions specific to its occurrence.[4] The Description module include tables that specify structural type, rock composition, geochronologic age, stratigraphic age, and structural measurements.

  [In CORDLink version 6.0, the "Feature" classes, "Geometric Object" class and the "Description" class among others provide a mechanism to handle complex geoscience data types. Structural geology, metamorphism, and stratigraphic ages are specifically contained within the model. Paleontology is minimally handled through the "Rock Composition" object. Other objects such as "DescRelation" have a paleontological element.]

- **Minerals/hydrocarbon databases**

  Basic minerology is handled within the "Rock Composition" table as a "mineralogy_desc" attribute. There are no apparent tables or objects dealing specifically with hydrocarbons.

- **Geochemistry/geophysics databases (raster and vector)**

  Geochemistry and geophysics are not included in the current model.

  NADM v5.2 (and CORDLink v6.0) is structured to deal with simple image catalogs. Multi-band and multi-sample type data cannot be handled in this simple catalog entity.

- **Assessment reports and associated databases**

  NADM v5.2 provides a Text or Document object for the inclusion of reports and other textual material. A specific bibliographic database definition is not part of NADM v5.2.

- **Bore hole databases**

  Simple borehole core sample geological attributes such as lithology, age and rock unit classifications are included in the model but NADM v5.2 does not specifically address all elements of borehole data.

[4] Brodaric,B., M.Journeay, S.Talwar, and E. Boisvert. CordLink Digital Library. Geologic Map Data Model Version 5.2. June 18, 1999.

3.2.2 Software independence and availability of support

- The NADM v5.2 data model is a logical model and is platform independent.
- The current CORDLink implementation is a prototype but is not a complete test of the scalability of the model. A true test is probably not possible given the implementation software platform.
- Members of the committee can view the html version of NADM v5.2 at the NRI3 site (http://nri3.cfs.nrcan.gc.ca). Members should contact Rick Morrison ((250) 363-0772 or rmorriso@nrcan.gc.ca) for a password.

3.3 POSC

POSC (Petrotechnical Open Software Corporation) is an international not-for-profit membership corporation. POSC's initial contribution was a set of specifications in the form of an industry recognized data model called Epicentre. The purpose of the model is to provide a mechanism to improve sharing of technical information within and among oil companies, oil field service companies, governments and other players in the industry.

3.3.1 Analysis of the POSC Candidate Model

3.3.1.1 Overview

Epicentre v2.2 is an extremely complex and extensive model. A critical point to understand about the Epicentre data model is that it is a logical data model, and as such, it is not directly implementable as a physical database.[5] Epicentre is documented precisely in the EXPRESS language, but EXPRESS is not the equivalent of a data definition language (DDL), such as SQL DDL. To build a POSC data store, it is necessary to transform Epicentre EXPRESS into a set of DDL statements using rules consistent with the target data store's database management software. POSC refers to this process as projection.

Although Epicentre is not directly implementable, it is an excellent entity source reference. Most geoscientific areas have been addressed to some degree by the model and so it can provide a good starting point for other design efforts.

3.3.1.2 Detailed evaluation

- **Geological maps (GIS and attribute databases)**

Epicentre supports the representation of simple and composite mappable geospatial

[5] POSC. (1997) About the Epicentre Data Model. http://www.posc.org/. Copyright 1997 POSC.

information using the "Spatial Model Topology Hierarchy." This is a traditional georelational model as implemented by various proprietary GIS vendors. The concept of "Properties" as well as "Earth Model Objects" also form part of Epicentre.

Geological processes, geochronology, and an extensive array of "Geological Objects" (e.g. faults, structural features, stratigraphic features) are defined within Epicentre. These definitions could be used as sources for the development of a cohesive model.

- **Minerals/hydrocarbon databases**

  Epicentre was developed with an industrial focus on exploration and production. The hydrocarbon side of the data model is extremely well developed. Mineral definitions are relatively rudimentary.

- **Geochemistry/geophysics databases (raster and vector)**

  Within Epicentre there is extensive coverage of geochemical attributes as related to oil and gas production. Epicentre contains an entity called "Analysis" which includes, among others, the subtype "Surface Geochemical Surveys", and a large variety of "Material Processing" subtypes. Non-petroleum industry geochemical analyses are not within the domain of Epicentre.

  Seismic data are extensively defined and exhaustively described within Epicentre. From "Seismic Geometry" and "Seismic Grids" through to "Seismic Earth Model Interpretations." Additional geophysical down hole and lab measurements made on core samples are also contained in the model.

- **Assessment reports and associated databases**

  Assessment reports and others such as "Incident Reporting" are handled by Epicentre. However, Epicentre does not support a bibliographic database system directly.

- **Bore hole databases**

  Epicentre has extensive petroleum borehole coverage. Entities include "Wellbore Boreholes and Volumes", "Well and Wellbore", Wellbore Position" and "Wellbore Devices" among others.

3.3.2 Software independence and availability of support

- Epicentre version 2.2 is publicly available via the web. See http://www.posc.org/.
- A mandate of POSC was to develop a system that enabled interoperability.
- POSC has built a projection tool for relational database management systems (RDBMS) and has used it to project Epicentre Version 2.2 into four types of RDBMS DDL: ANSI, Informix, Oracle7 and Sybase Release 4. Spatial Representation.

- POSC offers support for members and there is also a consulting division of POSC available which supplies expertise in the implementation of their specifications (see http://www.posc.org "Services" for more details).

3.4 ODP

The Ocean Drilling Program (ODP) is an international partnership of scientists and research institutions organized to explore the evolution and structure of the earth. To support its research activities ODP developed JANUS, an Oracle relational database. The database currently contains over 300 tables of ODP's marine geoscience data that are collected onboard the drillship JOIDES Resolution. The database includes paleontological, lithostratigraphic, chemical, physical, sedimentological, and geophysical data for ocean sediments and hard rocks.[6]

3.4.1 Analysis of the ODP Candidate Model

3.4.1.1 Overview

Like PPDM and POSC's Epicentre, JANUS has a strong borehole focus. However, JANUS is science based rather than industry based and thus expands on areas such as paleontology and sedimentology beyond that found in PPDM or Epicentre.

A unique feature of the JANUS model as compared with the other three models reviewed here is that it very clearly differentiates between interpretative data and measured data. For example, a radiometric age is regarded as measured data while a geological time scale designation such as "Jurassic" is considered interpretive. The model accomplishes this distinction by making extensive use of "concepts" to refer to interpretative data. In the JANUS context, a "concept" must be supported by a reference. As an example, the geological age "Jurassic" may be a "concept" *sensu* a GSC time scale[7] or *sensu* the Harland[8] time scale. In this way, the precise meaning of specified geological age may be identified and contributors of data to an implemented system are not restricted to a single, specified geological time scale.

The paleontology model was the first model developed within JANUS. Taxon designations, zones, and datums are all supported by "concepts" and references. In addition, all interpretations (e.g. a fossil identification) must also include a reference to the person or persons who made the interpretation.

3.4.1.2 Detailed evaluation

---

[6] Ocean Drilling Program Janus Web. http://www-odp.tamu.edu/database/. Accessed March 31, 2000.

[7] Okulitch, A.V. General Co-ordinator (1995) Geological Time Chart. The National Earth Science Series Geological Atlas.

[8] Harland,W.B., R.L.Armstrong, A.V.Cox, L.E.Craig, A.G.Smith & D.G.Smith (1990) A Geologic Time Scale 1989, Cambridge University Press, Cambridge.

- **Geological maps (GIS and attribute databases)**

  Simple positional data such as latitudes and longitudes are recorded in the “Hole” entity. The “Leg” entity records area descriptions, total miles transited and total miles surveed along with some additional speed data. “Depth_maps”, “Map_types”, and “Subsections” record various interval data. Time lines are also included with much of the positional data. The data model, as defined in March of 1996, does not specifically allow for complex spatial objects.

  JANUS has an extensive paleontological component. Fundamental to this component is the separation of interpretive from measured data. Basic data are recorded about a paleontological sample (e.g. fossil group, preservation, paleoenvironment, geologic age and zone) at time of collection and then these data may be reinterpreted in the lab as required. All identifications, be they of taxa, ages, zones or environments, must be supported by a referenced concept and assigned to the scientist making the identification.

- **Minerals/hydrocarbon databases**

  The JANUS model does not specifically include either mineral or hydrocarbon attributes. Any coverage is the result of inclusion in sedimentological or petrological entities.

- **Geochemistry/geophysics databases (raster and vector)**

  There is extensive coverage of geochemical attributes within the JANUS model. “Residues” and “Residue_Treatments” entities along with “IW_Test” (for interstitial water). Also “Gas_Chromatography_Sample”, “Rock_Eval_Sample”, and “Moisture_Density_Sample” entities among others.

  Geophysical requirements are also well represented in the JANUS model but they primarily focus on borehole related geophysics and not surface surveys. Both down hole and lab measurements on core samples are included. For example, natural gamma ray data, seismic logging data, p-wave loggers, gamma ray porosity estimators plus resistivity form part of the model. Raster data are not specifically handled by the model.

- **Assessment reports and associated databases**

  Assessment reports and other reference materials are not specifically included within the JANUS model but may be linked through bibliographic keys.

- **Bore hole databases**

  The JANUS model has extensive borehole coverage which is not surprising given the nature of the work of the Ocean Drilling Programme. Entities include “Casing”, “Cement”, “Hole_History”, “Bit_Manufacturer” among others.

3.4.2 Software independence and availability of support

- The JANUS model may be implemented on any relational platform.
- JANUS Web provides access to ODP's Oracle relational database JANUS. Non-proprietary data such as ODP & DSDP site information are available to the public.
- The data model and accompanying software are not publicly available.

4 Summary

Each of the four major models reviewed above has strengths and weaknesses (see Table I) when considered as candidates for the Canadian Geoscience Data Model. None of the models reviewed provide a common geoscientific model which will enable integration of data holdings and information delivery.

Requirement	NADM (NADM v5.2)	PPDM	POSC (Epicentre)	ODP (JANUS)
Focus	Science	Industry	Industry	Science
Spatial data (GIS)	Good	Partial (spatial extension in beta)	Good	Partial
Geological attributes	Partial	Partial	Partial-Good	Partial
Minerals	Minimal	Minimal	Partial	Minimal
Hydrocarbons	None	Good	Good	Minimal
Geochemistry	None	Partial	Partial - Good	Good
Geophysics	None	Partial-Good	Good	Partial-Good
Assessment reports and other documents	Good	Minimal	Good	Good
Borehole	None	Good	Good	Good
Model publicly available	No (password protected web site)	No	Yes	No
Software available	Yes (proprietary)	Yes (proprietary)	Yes (proprietary)	Yes (proprietary)
Support available?	No	Yes	Yes	Yes

Table I. Requirements as supported by reviewed models.

Both PPDM and POSC Epicentre are petroleum industry developed models. The PPDM model is well established with a very directed petroleum industry focus. Significant work would be required to extend this model into more scientific areas.

POSC Epicentre also has a strong industry focus but the model extends beyond basic industry needs through the addition of detailed scientific components in the areas of geological attributes, geochemistry and geophysics. Epicentre is unquestionably the most fully specified of all the models reviewed but it is also the most complex and thus 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. However, Epicentre can serve as an excellent entity source reference for other design efforts.

The NADM v5.2 data model and the ODP JANUS model are more science based than PPDM and Epicentre. NADM v5.2 has a focus towards geological maps and interpretive data. Unlike ODP JANUS, NADM v5.2 does not clearly distinguish between interpretive data and source or measured data. The vector representation within NADM v5.2 is especially well documented, and the geoscientific model approach is an essential element. However, NADM v5.2 is limited in scope so it is difficult to determine if it will be able to effectively handle disparate disciplines such as paleontology, geochemistry and geophysics among others.

The ODP JANUS data model excels in the areas in which NADM v5.2 is weak. JANUS very clearly separates interpretive attributes from source or measured ones. This is particularly important if the data are being collected in an environment that is not dominated by a single set of standards. JANUS also includes paleontological, geochemical and geophysical specifications along with an extensive borehole coverage. The JANUS model does not, however, specifically include geoscience models or map data to the extent of NADM v5.2.

The Canadian Geoscience Data Model needs to provide a balance between an industry focus and a scientific one. None of the four key candidate models accomplishes this although POSC Epicentre comes closer than the other three. If the Canadian Geoscience Data Model was to be derived from the four key candidate models reviewed, it would do well to:

- use the spatial and mapping components from NADM v5.2,
- use the clearly defined approach to distinguish between interpretive and source data from JANUS,
- draw upon each model to define geological attributes (e.g. structural geology from NADM v5.2, paleontology from JANUS),
- use the industry based models for specifying hydrocarbons,
- extract relevant geochemical, geophysical, and borehole components from PPDM, Epicentre and JANUS as appropriate,
- handle report and other document type data as specified in either NADM v5.2, Epicentre or JANUS or perhaps a combination of these,
- find a better specification for minerals data, and finally
- establish a single unifying approach that would expand upon the best elements of each model.

As noted in the "Overview of Objectives and Scope of Model Review" section, the candidate models reviewed above are but four of numerous possible models worth considering. During the survey phase of this project, we will be gathering further information on these and other models as put forward by the stakeholders.