METADATA + METAINFORMATION = METACONTENT: UNITING THEORY AND PRACTICE USING ORACLE DESIGNER

Henry Kucera, Mercator Systems &
Andrew Faulkner, Ministry of Environment, Lands & Parks, British Columbia

INTRODUCTION

Advances in technology have so improved the efficiency of data collection that organizations are finding it difficult to keep up with the growing flood of data. Many organizations continue to use the data for single narrowly focused applications in which only the local creators of particular databases may be very knowledgeable about their characteristics and legacy. Unfortunately, while local analysts may use a particular data set with confidence, consideration is seldom given to recording the history of the database in a form that can be conveyed to *foreign* users. By foreign user we mean those who had no part in collecting or storing the data originally and are separated from the data collection activity by:

- time - viewing historical data at some point in the future;
- space - physically separated or having limited proximal access to the data and data collectors;
- subject area - not being a specialist in the information domain that the data represents.

Foreign analysts who wish to query and/or process the data, are equally in need of information about the assumptions made prior to collecting the data, the sampling procedure, potential sources of bias, inconsistencies and/or error. Lack of knowledge of these factors can limit the manner in which particular data can be manipulated and interpreted. Because of this, a wealth of information collected at enormous expense remains locked in databases and cannot be mined to extract additional knowledge. A key to transforming data to information and ultimately to knowledge is *Metacontent.*

METACONTENT

Metacontent is data that describes both database organization and the data's fitness for use. Database organization involves:

- physical storage
- and schema.
- Fitness for use involves:
- content,
- ownership,
- collection and processing methods,
- lineage,
- resolution,
- accuracy,
- spatial location and extent,
- temporal location and extent.

Metacontent is often referred to simply as metadata. Unfortunately, usage of the term metadata has become ambiguous and means different things in different contexts. The information management community reacted to the ambiguity of

the use of the term metadata by introducing the additional terms metacontent and meta-information, and refining the use of the three terms (Podehl 1993, Sundgren 1993). Metacontent is the generic term; metadata and meta-information are the two categories of metacontent that are related as shown in Figure 1.

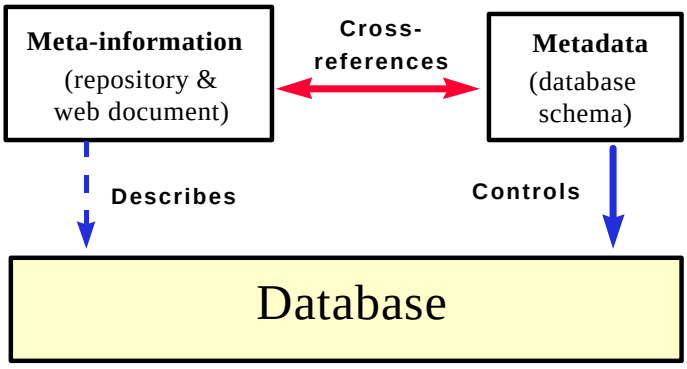

Figure 1 Metacontent: relationship between data, metadata and meta-information

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). Metadata normally 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 (Kucera and Flaherty, 1997) Meta-information (description and/or 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.

How important is it to understand meta-information? We use meta-information every time we frame a question. As an example, lets investigate if a community may be 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. She also must determine the groundwater characteristics of the underlying geology. To begin the analysis, she places parameters on age of information, source of information, accuracy, and completeness. Is the information she needs in paper (maps, photos or reports) or digital form (LANDSAT, well logs, or seismic surveys)? Does she 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 likely scenario? A recent non-fiction national bestseller " A Civil Action" (Harr, 1996) which chronicles a personal injury case against major US manufacturing companies is based on exactly this type of investigation.

It makes little sense to expend resources collecting data that cannot be used and re-used for a variety of purposes. The best way to ensure that data are re-usable is to describe their provenance and storage method in sufficient detail that foreign users are able to assess the data's fitness for an intended purpose, and achieve access to the data without undue difficulty. The importance of metacontent grows as the heterogeneity of data sources increases, and as organizations increasingly plan to retain data for posterity.

Figure 2 illustrates the roles by which metacontent can support user interaction with a data store. A metacontent repository assists in 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; and it provides meta-information for browsing and integration during decision support. Not all metacontent repositories serve all these roles; a repository designer may consciously choose a subset of roles to serve.

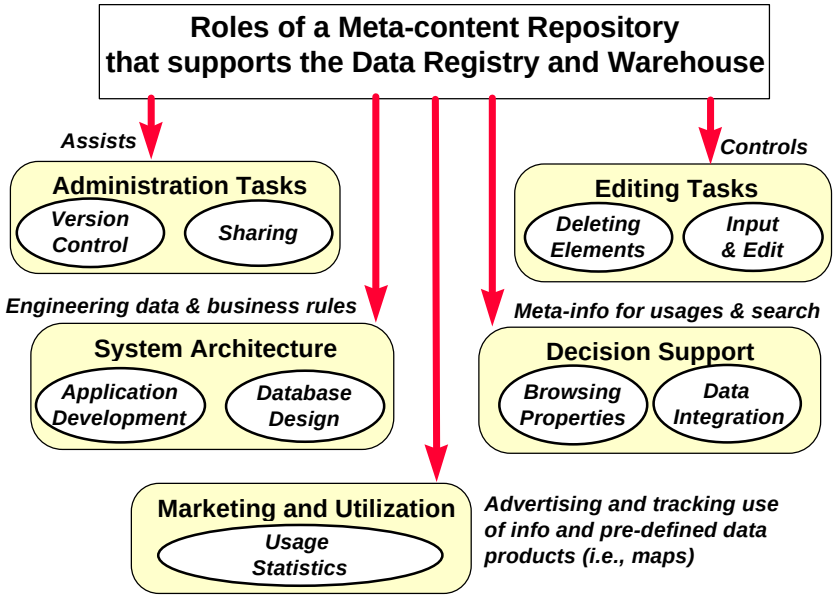

Figure 2 Roles of a metacontent repository. (MELP, 1997)

MELP Data Registry System

In British Columbia, at the Ministry of Environment, Lands & Parks (MELP) the Data Registry System is the key to describing the data warehouse contents. Once content is known, there also must be a way to query and retrieve data from the warehouse. The Data Registry can provide the means to intelligently search for and retrieve information. This section will describe the current system for data warehousing and metacontent management at MELP.

Data Warehouse Architecture

An effective data warehouse strategy can change the information utilization landscape. A data warehouse is a collection of subject-oriented data of known quality, that is low in volatility and may be time-variant. The warehouse may include the basic tools to access and extract information. It is designed to be accessed by multiple users to access information that previously was not available using single purpose operational database systems.

A primary goal of the Information Systems Branch (ISB) is to enable the migration MELP data from the central administrative computer systems to the distributed desktops of the ministry's decision-makers. ISB has addressed these goals by creating the MELP data warehouse. The warehouse contains read-only copies of data restructured for query and analysis, and stored in standard formats in a single database. Automated processes copy the information to regional servers, where it is used by operational staff. MELP's data warehouse has the following logical components:

- A central tabular database, consisting of tables and views stored in a relational DBMS.
- A central spatial data store, consisting of map layers and features stored in a GIS, in disk files.
- A WWW site, providing access to parts of the catalog and databases by the public and other agencies.
- Copies of data from the central databases at each regional site, replicated there to give faster access to local users.
- A central data registry, or catalog, listing the spatial and attribute data sets, and their metadata.
- Viewing and query tools.

DATA REGISTRY ARCHITECTURE

The intent of the Data Registry is to give all operational groups in the ministry easy access to the same information, thereby encouraging communication and cooperation between program areas. As the Registry becomes more accessible, it also will facilitate joint resource management for cross-ministry environmental planning and regulation. It currently contains over 1100 information sources with reference to over 230 GIS themes. The general public also has access to a subset of the data through the internet.

The Data Registry is implemented in an Oracle database located on the HP UNIX servers at ISB in Victoria. The Data Registry Version 3.2 consists of linked tables that can be updated and queried with an Oracle Forms user interface as well as a web query interface.

The creation of the Data Registry has been an organic process. The completeness of the content and the structure of the data model have grown and changed as the supporting technology has improved and the understanding of the information resources have increased. The design and content of the present Registry Version 3.2 was defined by users interacting with ISB staff. Figure 3 shows the context of the current system.

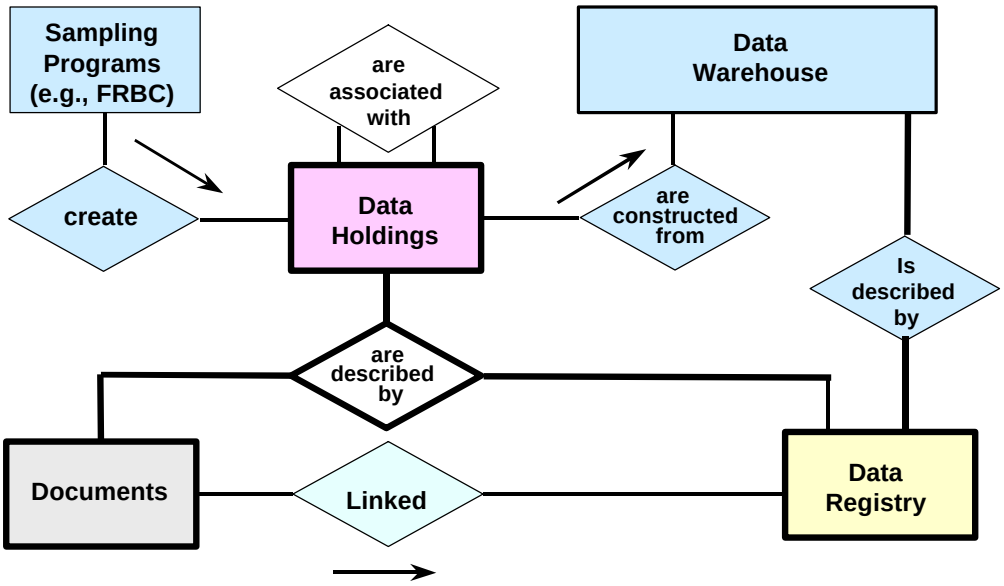

Figure 3 MELP data warehouse and Data Registry system context. (MELP, 1997)

MELP is responsible for: administering programs which collect data through sampling, monitoring or enforcement, creating and managing the data holdings which result from the programs, creating and administering a data warehouse that provides data to a large distributed set of users. The focus of the Data Registry is on describing what the information source was, what form the data is currently in, and who the contact person is for it either getting the data or more information about the data. The core data is illustrated in a simple entity relational diagram in Figure 4.

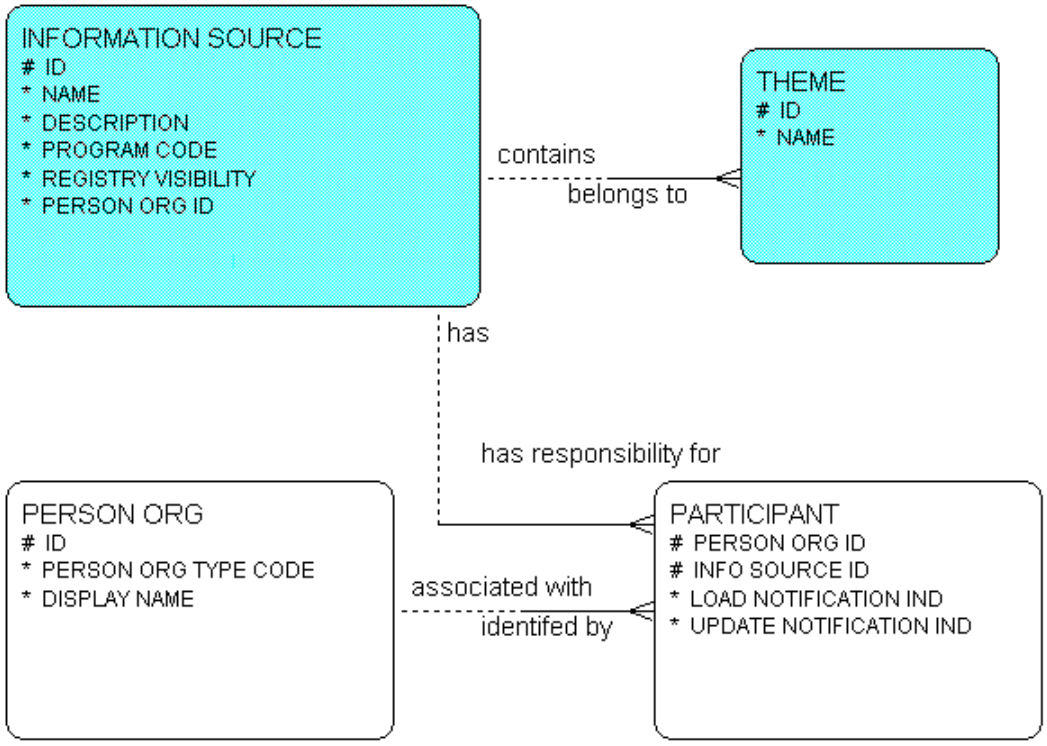

Figure 4 Data Registry main entities

The key entity is the *Information Source* and associated to it are *Person/Org*anizations which have a specific role in the association such as data custodian or primary contact. There may also be in association with an *Information Source* a GIS coverage which is noted in *Theme.*

The technology used for the Data Registry has been exclusively Oracle. The current deployed configuration is:

- Oracle 7.3.2.3.0
- Oracle Forms 4.5
- Oracle Designer 1.3.2
- Oracle Web Server 2.1 & 3.0
- Cherry Pie (Oracle Corporation, Netherlands, 1997)
- Oracle ConText Version 1.1.2.0.0

The system is depicted in Figure 5. For the purposes of this paper the Data Registry System will be discussed in terms of its management of data warehouse holdings but in can (and does) contain information sources that are available through other access mechanisms such as hard copy reports, tape, diskette, ftp site, and operational data stores.

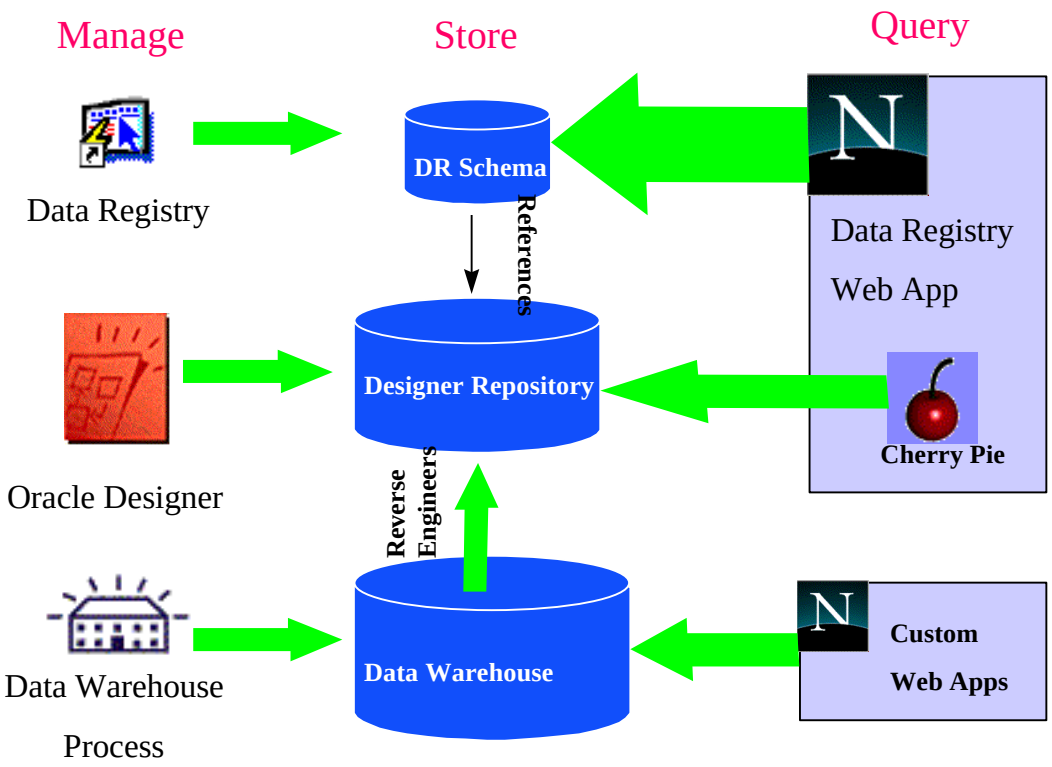

Figure 5. MELP Data Registry Architecture

The metadata management is done with Oracle Designer and focused primarily on the holdings in the Ministry's data warehouse through the integration between the web, Oracle Designer Repository, ConText and Oracle Forms custom application.

An *Information Source* is entered into the system through the Data Registry Oracle Forms 4.5 application. Information sources stored in the data warehouse have the schema linked by inclusion of the web URL (universal resource locator) that, when invoked from a web browser, executes the Cherry Pie (Oracle Netherlands, 1977) application. This provides definitions of entities and domains from the Oracle Designer repository. The repository is populated by reverse engineering the data schema from the data warehouse. Figure 6 shows the Information Source Summary screen for the Oracle Forms 4.5 application. The blue check marks beside the buttons indicate that further metacontent is available for those categories of information. The categories are:

- People,
- Instances,
- Applications,
- Map Themes,
- Associated Map,
- Keywords,
- Comments,
- and Information Classification.

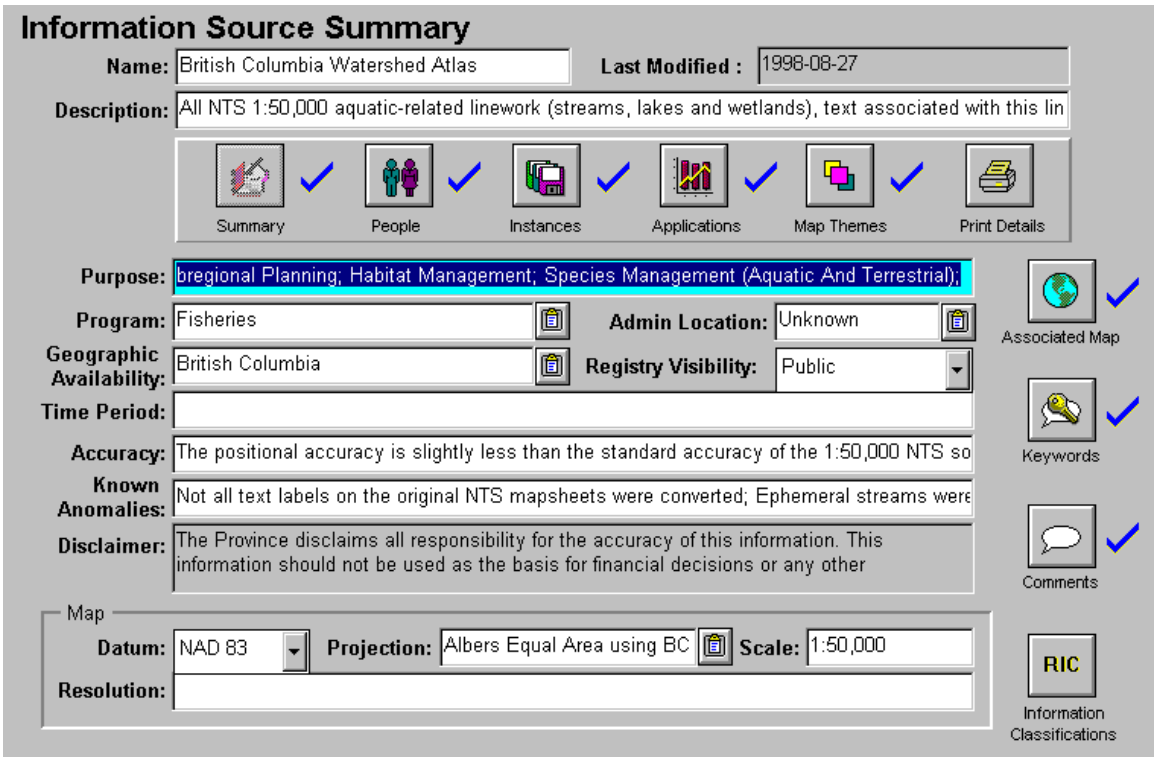

Figure 6 Data Registry information source summary screen

There are a number of custom web applications that provide government and public users direct access to information holdings in the data warehouse. These may also be linked as a referenced application by inclusion of the web URL in the *Information Source* entry.

WEB INTERFACE

The internet and intranet users can search the data holdings using a simple web interface to locate the meta-information (who, where, description, purpose, currency) and drill down to metadata (instance, table, column, map theme).

http://www.elp.gov.bc.ca/apps/dr/

The functionality is focused around providing enough information for the user to determine:

- Is there data relevant to my topic?
- Is the data available appropriate for my needs?
- Where is the data located?
- Who is the responsible for the data?

Oracle ConText (version 1.1.2.0.0) provides powerful search by key words across the INFO_SRCES.TITLE and the INFOR_SRCES.DESCRIPTION fields. ConText provides full-text retrieval and advanced linguistic services, tightly integrated with the Oracle relational database. Implementation of this technology is crucial to the systems ability to satisfy user queries where the thesaurus of possible search terms is unbounded. Since many different data administrators are including many different types of information it was found to be impractical to require specific “key wording” to be associated with each information source record. Instead the data administrators are asked to enter a title for the information source that is definitive for their subject specialization and then to enter as complete a description as possible so that others may understand the content. Word searches from the web may use enabled to use either or both text fields during a query.

US STANDARDS FOR META-CONTENT

The primary objective in standardizing the metacontent for the government information resources is to facilitate user access to this information as envisaged by various single window scenarios. A single view of the available government information resources is presented to the user. This view would hide the technicalities of information retrieval, indexing, display and related characteristics of individual systems. The user would have a single consistent view although they would be getting data from diverse sources. Some standards that are currently being developed for metacontent in the United States include the Government Information Locator Service (GILS) and the FGDC Content Standard for Digital Geospatial Metadata.

GOVERNMENT METACONTENT STANDARDS: GILS

The Government Information Locator Service (GILS) is a computer platform independent system for locating government information in a decentralized collection of databases. GILS systems or locators are made up of searchable databases of GILS records which indicate what information is available, where it is located and how it may be accessed or acquired. A GILS record is not the information itself, but a standards-compliant description and a pointer to a media-independent information resource. GILS records can describe a collection, a service, a system, a web site, a publication or an individual electronic document. They can contain a direct link (Uniform Resource Locator or URL) to a networked information resource. They can also describe how to obtain information that is not available on via an electronic network such as the Internet or a departmental Intranet.

GILS originated in the United States and U.S. federal government agencies were required by law to implement this government-wide service beginning January 1996. This service provides users with a means of finding government information, located in locate and remote systems. Users find information by formulating system independent at the desktop and transmitting these to a remote database containing GILS records. The queries are presented to the remote database using an international protocol for information retrieval called Z39.50. Z39.50 defines a common system interface that allows users to search one or more databases and to receive a consolidated set of responses to each search query. Recognizing, that Z39.50 implementations will not be prevalent at the user's desktop, government information providers typically provide a gateway to this information service. These gateways include support for the Internet-HTTP standard, supported by web browsers, and thereby provide access to government information for anyone who can access the Internet.

METACONTENT FOR SPATIAL INFORMATION: FGDC

The *Content Standard for Digital Geospatial Metadata* was created and is administered in the US by the Federal Geographic Data Committee (FGDC 1994). The FGDC was established to promote the coordinated development, use, sharing, and dissemination of geographic data. The FGDC has representatives from the US Departments of Agriculture, Commerce, Defense, Energy, Housing and Urban Development, the Interior, State, and Transportation; the Environmental Protection Agency; the Federal Emergency Management Agency; the Library of Congress; the National Aeronautics and Space Administration; the National Archives and Records Administration; and the Tennessee Valley Authority. Additional Federal agencies participate on FGDC subcommittees and working groups. The Department of the Interior chairs the committee. FGDC subcommittees establish and implement standards for data content, quality, and transfer; encourage the exchange of information and the transfer of data; and organize the collection of geographic data to reduce duplication of effort.

FGDC participated in the plenary and working meetings of ISO TC211, which were held January 20-24, 1998 in Sydney, Australia. TC211 Working Group 3 is developing a standard for geographic data content, with which the FGDC plans to comply when it becomes final in 1999.[1]

[1] FGDC Newsletter, March 1997. Published by FGDC, US Geological Survey, Reston, Virginia.

OTHER INITIATIVES IN SPATIAL DATA REGISTRATION AND DISTRIBUTION

Other jurisdictions are beginning to put together projects or programs to deliver distributed warehouse-type capability. These efforts are called National Clearinghouses, National Spatial Data Infrastructures, or Frameworks. The small sample of ongoing major projects gives some idea of the importance that has been given to delivering tightly integrated metacontent repositories and accessible digital warehouses.

INTERNATIONAL EFFORTS IN METACONTENT STANDARDAZATION: ISO TC211

Formed in 1994, ISO/TC 211 is an international effort to establish standards in geomatics (geomatics deals with the collection, interpretation and management of geographically referenced information). TC211 is working on standards that specify methods, and services to manage, acquire, treat, analyze, present, and transfer geographically referenced information in digital format between different users, systems, and sites (Kucera and O'Brien, 1997).

The objective of the standard is to provide a procedure for describing digital geographic datasets so that users will be able to locate and access geographic data and determine whether the data in a holding will be of use to them. By establishing a common set of metadata terminology, definitions, and extension procedures, this standard will promote the proper use and effective retrieval of geographic data. Supplementary benefits of this standard for metadata are to facilitate the organization and management of geographic data and to provide information about an organization's dataset to others. This standard for the documentation of data furnishes data producers the appropriate information for them to characterize their geographic data and it makes possible dataset cataloguing enabling data discovery, retrieval and reuse.

US NSDI INITIATIVES

The responsibility for coordination of the NSDI development has been given to the Federal Geographic Data Committee (FGDC) by Executive Order from President Clinton in 1994. The NSDI concepts were initially developed by the Mapping Sciences Committee (MSC) of the National Research Council and published within a report entitled, "*Towards a Coordinated Spatial Data Infrastructure*". The MSC produced a second report in 1994 entitled, *"Promoting the National Spatial Data Infrastructure through Partnerships"* which identified several key elements that should be common to future partnerships:

One section of the report outlines the concept of the National Geospatial Data Clearinghouse, which is defined as " a distributed network of geographic producers, managers and users linked electronically". Other sections set out the procedures to be followed with respect to data standards and describe the need to develop a National Digital Geospatial Data Framework.

To deliver the NSDI the FGDC has established a Competitive Cooperative Agreement Program to help form partnerships to facilitate the development. This program provides limited funding for cooperative projects with State and local government agencies, institutions of higher educations and private organizations. The first nine grants were made in 1994, mainly to promote the implementation of State nodes for the National Geospacial Data Clearinghouse.

FGDC has defined the core geospatial framework as a "basic, consistent set of digital geospatial data and supporting services that will:

- provide a geospatial foundation to which an organization can add detail and attach attribute information.
- provide a base on which an organization can accurately register and compile other themes of data, such as soils, or geology.
- orient and link the results of an application to the landscape.

The information content of the framework includes seven elements: geodetic control, digital ortho-imagery, elevation data, transportation data, hydrography, administrative boundaries and cadastral information.

AUSTRALIA - NEW ZEALAND NSDI INITIATIVES

The Australia New Zealand Land Information Council (ANZLIC) is the key inter-governmental council responsible for co-ordination of land and geographic information management in Australia and New Zealand. ANZLIC provides focus and leadership for the land and geographic information community in both countries. It was originally

established in 1986 as the Australian Land Information Council (ALIC) by the Australian Prime Minister and State Premiers in response to a "clear and growing need to co-ordinate the collection and transfer of land-related information between the different levels of government; and promote the use of that information in decision-making."

- the primary objective of a national data infrastructure is to ensure that users of land and geographic data who require a national coverage, will be able to acquire complete and consistent data sets meeting their requirements, even though data is collected and maintained by different jurisdictions.
- it envisages a distributed network of data bases linked by common standards and protocols to ensure compatibility, each managed by appropriate custodians nominated in each jurisdiction. ANZLIC will not create the infrastructure itself, but will assist the community to define and describe it 'as a coherent national entity'.

The Australian/ New Zealand NSDI proposal is similar in scope to the efforts documented in this paper but are deployed at a national rather than provincial scale.

EUROPEAN SDI INITIATIVES

During the last 18 months, the EU Directorate General DGXIIIE, EUROGI and interested parties in the European GIS community have prepared a document entitled *GI2000: Towards a European Policy Framework for Geographic Information.* This document has served to "open up the discussion concerning the need to establish a European policy framework for geographic information through which data can be created, marketed, used, reused and shared in a cost-effective manner for the benefit of society. The document points out the potential of a good GI policy for providing better and more efficient government as well as new business opportunities for the nascent European geographic information industry."

At the professional and inter-agency level there have been recent moves "towards a European approach to harmonizing activities concerning spatial data." Examples cited included MEGRIN - the Multipurpose European Ground Related Information Network, set up as a collaborative project by CERCO (Comité Européen des Responsables de la Cartographie Officielle) in 1993 and EUROGI - the European Umbrella Organization for Geographic Information, set up in 1994.

CEONET: A NODE ON THE CANADIAN GEOSPATIAL DATA INFRASTRUCTURE

The Canadian Earth Observation Network (CEONet) is an initiative by the Canadian government to create a national infrastructure for providing access to earth observation archives and other complementary spatial databases. This initiative is being driven by the requirements of Canadian users for better access to earth observation data, and by the opportunities for Canadian industry afforded by the rapid growth of the international market for earth observation data, services, and network systems.

CEONet provides a clearinghouse for suppliers and users of geospatial data and related services and is a one part of the Canadian Geospatial Data Infrastructure (CGDI). The CGDI is a government-driven initiative to promote and develop a common means for interoperable information (St. Laurent, 1997). The intent of the CGDI is to define a framework that brings together data providers, distributors, technology providers and users of information. The CGDI is similar to other national initiatives that will integrate with the emerging Global Geospatial Information Infrastructure as it links Canada to international efforts

CEONet contains information on Canadian and International geomatics and Earth observation organisations, products and services. CEONet is a World Wide Web (WWW) based system which:

- provides a single point of access for both consumers and suppliers of data.
- Uses the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), for both collection-level and product-level metadata.
- offers consumers a straightforward interface with which they can search for data that may be distributed among many disparate remote sites.
- Uses the ANSI/NISO Z39.50 search protocol as the means of supplier connectivity, and.
- uses Isite, a public-domain software package which supports Z39.50 and CSDGM, to handle searching.

FUTURE DEVELOPMENT OF THE MELP DATA REGISTRY SYSTEM

The Data Registry System is just beginning to contribute to the work of MELP and enhance its communication and dissemination of corporate data to the wider community. In recognition of the influence that US standards of metacontent will have on the tools marketplace several developments are strongly indicated that will allow the ministry to leverage the accessibility of its current data holdings. Some of these directions are:

- improved data model to become current with TC211;
- release of Data Registry 4.0 to include this level of detail;
- development of Z39.50 data views;
- and experimentation with metacontent exchange between repositories.

Access to attribute data for decision support within the data warehouse involves the deployment of business intelligence tools that themselves retain their own metacontent. MELP is currently deploying Oracle Discoverer 3 which imposes the overhead to manage its end user layer (EUL). Although there is some connectivity provided between Oracle Designer repository and the Discoverer EUL, and the promise of tighter coupling in the future, the following activities are contemplated in the short term:

- development of standards for Discoverer "Business Area" creation;
- and implementation of EUL query functionality as developed for the Designer Repository with Cherry Pie.

CONCLUSION

Building a data warehouse can be very difficult. Adding spatial data to the content increases the complexity and length of the project as well as adding demanding metacontent management requirements. The data warehouse architecture must be founded on modern, open data access standards. Its pay-off if taken at face value, may not appear worth the effort. However, with a sound design it will quickly become apparent that the primary reasons a data warehouse is built will only be a minor considerations once other value-added applications present themselves (Kucera and Rawlings, 1997). The key is to implement a primary data warehouse infrastructure that can:

- provide easy, on-line access to a wealth of information that has been gathered at great cost;
- provide industry, educational institutions, researchers and government agencies with the basic data from which to build valuable information products and applications;
- stimulate the growth of companies that exploit such a data resource;
- improve analysis and decision making;
- improve capability in detection and response to hazards and emergencies.

The architecture presented in this document provides a data warehouse infrastructure that combines:

- text, image and spatial data support in production databases;
- applications development tools for client/server and Inter/Intranet;
- information repository;
- inter/Intranet communications capability.

This combination means that the existing investments in technology are protected and reused while new pieces can be added from a larger technology toolbox. Challenges of today and tomorrow demand comprehensive vision and an understanding of how to construct an end-to-end solution from the technology toolbox.

ACKNOWLEDGMENTS

The authors wish to acknowledge the following contributors to development of the Data Registry system: Pangaea Systems Inc., LGS Systems Inc., the MELP Information Management Working Group, and in particular Denise Owen, Richard Lay, and Heather Lehman.

REFERENCES

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

Kucera, H.A., and O'Brien, C.D., (1997), Harmonization of the SQL/MM Spatial and TC211 Standards, Joint Canadian Contribution to ISO/TC 211 and JTC1 SC21/WG3, CAC SC21 WG3 N429, Madrid, January.

Kucera, H.A., and Rawlings, J., (1997) The Mercator Alliance: A Spatiotemporal Framework for Multidisciplinary Systems, 11th Annual Symposium on Geographic Information Systems, Feb., Vancouver, British Columbia.

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

MELP: Data Warehouse Meta-content, URL http://www.elp.gov.bc.ca/isb/projects/meta/metada.htm, April 1997.

Oracle Netherlands (1997), Cherry Pie, URL http://www.elp.gov.bc.ca/apps/dr/cpie/cherry_pie.htm

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

St. Laurent, L., Kucera, H.A., and E. Keighan, (1997), Geospatial Data Warehousing: An Architecture Presented at GEOMATICS IN THE ERA OF RADARSAT, GER'97, May 25-30, 1997, Ottawa, Canada

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.