Integrating the Data Repository with the Publication Process

By Andrew Moore[1], Linda Guay[2], Ross Murray[1]

[1]Geological Survey of Canada, Natural Resources Canada
[2]ESSInfo, Natural Resources Canada601 Booth Street,
Ottawa, Ontario
Canada
K1A 0E8
Telephone: (613) 943-0776
Fax: (613) 947-9518
Email: amoore@nrcan.gc.ca

INTRODUCTION

In 1999, the Government of Canada made a commitment to become known "as the government most connected to its citizens, with Canadians able to access all government information and services on-line at the time and place of their choosing" (1999 Speech from the Throne). The resulting Government On-Line (GOL) initiative is the plan to deliver the Government's programs, services and information through the Internet. The Geological Survey of Canada (GSC) through its parent agency, Natural Resources Canada, is moving in a very significant way to support this mandate. As a result, the information technology / information management (IT/IM) and geoscience communities have felt the pressure to organize a Geoscience Data Repository (GDR) in which to store the wealth of the GSC's services, knowledge, information, and data in a fashion that makes it easily accessible by the public and, in particular, by decision makers. The GDR is a network of servers, applications, and databases that make up the corporate archive; this infrastructure requires standards that are easy to implement through the use of common practices and methods.

Beyond the basic infrastructure issues and requirements involved in defining a corporate IT/IM strategy, a greater level of coordination is required between management, scientists, and IT/IM staff. This coordination will facilitate the 'corporate' archiving and on line access to GSC's information themes. Geological mapping is one such theme in which a distributed database based on "version 5.x" of the North American Data Model (http://geology.usgs.gov/dm/) is being developed to archive not only the scientific interpretation and geospatial components, but also a normalized science language that describes the key attributes of the geologist's interpretations in the geologic map explanation, or legend. In order to facilitate the loading of information to the GDR, a process is being developed to integrate this information population into the GSC publication process. As with any science-based organization, the publication process is very important in ensuring a consistent and high quality product. By developing a process that efficiently manages the publication flow and integrates the information used to develop the publication, the GSC can meet the increasing demand for our knowledge and quality information while fulfilling our GOL mandate.

DRIVERS

The potential impact of geoscience information on both traditional and non-traditional decision makers is greatly increased when it is made available through the Internet. Even traditional publications such as maps, reports, and bulletins can take advantage of the wider distribution and easier access that the Internet provides. This is an obvious advantage to management, and there has been a push to develop these Internet portals. Unfortunately, what is less obvious is the amount of time, effort, and planning that is required to build the infrastructure to support these portals. Using traditional GSC science products such as a surficial geology map as an example and following it through the process to make it digitally accessible through the Internet, the obstacles that currently exist in getting this information to the corporate Geoscience Data Repository (GDR) are quickly identified. This is primarily the result of a traditional publication process thrust into the information management world. The knowledge and information that we want to make available through the Internet first has to go through the publication process. The publication process ensures quality information by providing several edit and peer review steps. This system works relatively well for traditional paper products; however, the observed primary information and data collected and compiled during the science project is now as much in demand as the published interpretation. The demands for high quality, consistent surficial geology information come from many applications such as groundwater protection, industrial mineral management, protected

lands, basic research, mineral exploration, engineering, and environmental assessment. Prior to digital collection and digital map compilation, such information resided in field books or on aerial photographs; therefore, access to the primary information was extremely difficult. Currently, the primary information is archived on an ad-hoc basis, as the formal process to manage it is only now being developed. In the surficial geology example, we wish to develop a science language, or taxonomy, that can be used to represent the key attributes of the map legend. The legend will remain the domain of the scientist in order to reflect his interpretation of the glacial history and geological story, but the science language will provide a standardized way of querying this knowledge.

In developing tools to allow geologists to 'parse' (extract a consistent terminology from common attributes) their legends so that the spatial objects (points, lines, and polygons), original legend, metadata, symbology, and science language can be archived in the GDR, we realized that much of this same information is required by different working groups in the publication process. Divisional management and Earth Sciences Sector Information (ESS-Info) Publications require a permission to publish (P2P) form; critical reviewers must also be able to access much of this information to review the science; Editorial and Cartography sections, the Book Store, and Earth Science Information Centre (ESIC) draw upon, modify, and add certain components of the publication / information set. The preparation of the digital information to be accessed through our GDR and the preparation of the interpretation for publication are not integrated in a common process. If we can coordinate these two activities, we can not only make the effort more efficient but we can also improve the quality of the information and streamline the publication process.

CURRENT SITUATION

Each of these administrative groups in the GSC has adapted to information technology, but each has developed a solution that fulfils its individual needs. Unfortunately these solutions are not integrated and so the process of preparing and publishing a map requires a significant amount of interaction with several individuals and groups. For example, a P2P form must be completed, which includes author, title, project, etc. A divisional archive may be used to track publications, and that database requires author, title, project, etc. As the publication moves to critical review, a form is completed that includes author, title, project, etc. Once through internal review, an Open File number is requested. This information is entered into a database and, again, it includes author, title, project, and so forth. This continues through several steps of the publication process, which requires several different data fields of information beyond author, title, and project. The re-entry of identical information at various steps of this process has resulted in different titles in different databases for the same publication, legend conflicts between Editorial and Cartography that can delay the release of the publication, and other inconsistencies that render the process less efficient and accurate than it can be.

There is currently a project involving the GSC and ESS-Info division to develop a distributed network of corporate servers to archive, through the GDR, the GSC's geologic information in consistent data models; the Cartography section is a registered International Organization for Standardization (ISO) shop that requires standard forms and procedures and is using a database to manage its production flow; the Publications section employs a database system to track and manage GSC's publications; the Book Store uses a database to manage its stock and to keep track of GSC's clients; and ESIC, formerly the GSC Library, manages a database also used as a key source in Internet metadata access. Science divisions are developing the means to archive the published information in standard data models and taxonomies, such as in the surficial geology project, that will reside on the GDR. GSC Projects and ESS-Info are currently populating the GDR with surficial and bedrock maps from Cartography that are formatted into the designed data model (NADM 5.x). The GSC is to develop tools (such as the legend parsing tool) to link this digital map information to the science language. Publications, Editorial, Cartography (all part of ESS-Info) and the GSC science divisions are currently reviewing the publication process, with the goal of producing an updated publication process that will adapt to the IM infrastructure being developed. This will produce a process map or diagram that identifies what is required at each stage of the publication process and the person with whom the responsibilities lie. In effect, it will embed the infrastructure work (the same parsing tool as above) into a corporate process (publication).

TECHNOLOGY

The technology (software and hardware) and expertise that currently exist in ESS-Info and GSC can support the integration of thematic database standards development and corporate processes development, and we are now working to build this system and implement it in the GSC. Figure 1 shows how such a process can support the requirements of the publication process while integrating the demands and inputs of the corporate archiving work. Starting in the upper left of the figure and, again using a surficial geology map as an example, the author logs in to the Intranet system and indicates that he/she will be submitting an Open File (OF) map for publication. A P2P form (fig. 1) is then displayed and the required information is entered by the author. This information is submitted through the form to GDR, indicated here as the "The

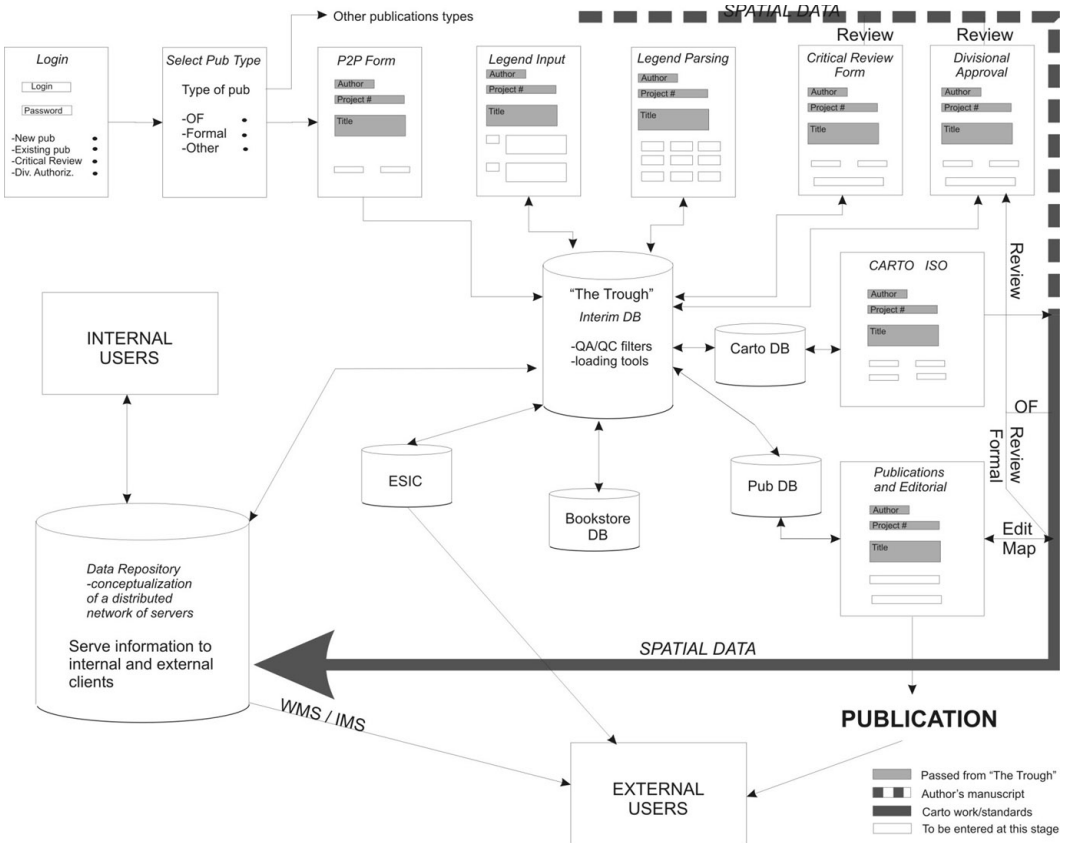

Figure 1. Planned publication flow, starting in the upper left part of the diagram and proceeding to the Geoscience Data Repository at lower left and publication at lower right.

Trough", for conceptual clarity and is simply a subset of the GDR that contains a variety of information needed for map preparation but flagged as unavailable for public consumption. In this process, the author ultimately uses the legend parsing tools to enter the legend description into the system and to parse the legend to the science language. As the publication and digital data move through the process, forms are displayed where many of the fields (such as the title) are already completed as the forms and/or local databases are being fed by the central database. The parsed information is also available for critical review in conjunction with the published product, providing a quality control step, and the database standard is then embedded in a corporate process.

Each step of the process triggers an action. For example, once the legend and P2Pform are submitted, an email is sent to division management indicating action is required on a publication. The sub-division chief reviews the P2P and assigns an internal critical reviewer. This triggers an email to the critical reviewer to review, and the process continues. Once Editorial and Cartography requirements are complete and divisional management has OK'd the publication (all electronically) the information is archived to the GDR, making it available to the public, and the publication is released. Each group or individual is assigned responsibility for each component of the information/publication set at each stage of the process.

REQUIREMENTS/RESULTS

The expertise to support this project is available in-house, but this project requires a significant amount of planning, cooperation, and effort. Each stakeholder (science divisions, Publications, Editorial, ESIC, and Cartography) must be consulted to ensure that each of their needs is met and that the responsibilities at each step of the process are clearly defined. For such a streamlined system to be effective, there must be 'buy in' from all groups involved. Consultation and participation by these groups is required in the planning and development of this system. Terrain Sciences Division has offered to pilot this process on behalf of the GSC science divisions. Implementing infrastructure standards of the IT/IM strategic plan will also support this project. For example, once the science language and data model for surficial geology are established, Terrain Sciences Division will not support a map publication until the legend has been parsed according to this corporate language. This will ensure that the publication process, and the archiving and access of our corporate knowledge are tightly linked.

A contract was awarded in December 2002 to build such a system based on commercial, off the shelf software (COTS) and custom modifications using standard tools and software. As of June2003, the initiative is in the final stages of user acceptance testing and will then be deployed as a pilot project using Terrain Sciences Division publications. Not only will the GSC be able to more efficiently produce publications and archive its vast information holdings, but the public will benefit by being able to access quality information and quality scientific interpretations in publications in a timely and coordinated fashion.

REFERENCES

Information Technology and Information Management Task Force, 2003, Earth Science Sector Information Management Plan: Earth Science Sector- Natural Resources Canada, internal document.