Natural Resources Canada
Ressources naturelles Canada

Effective project-level information management: development of a standardized protocol for capturing metadata within the Baffin Bay Basins Project of the GEM-Energy Program

P. Dhesi, J.W. Haggart, and T. Brent

Geological Survey of Canada

Technical Note 7

2014

Canada

Geological Survey of Canada
Technical Note 7

Effective project-level information management: development of a standardized protocol for capturing metadata within the Baffin Bay Basins Project of the GEM-Energy Program

P. Dhesi, J.W. Haggart, and T. Brent

2014

© Her Majesty the Queen in Right of Canada, as represented by the Minister of Natural Resources Canada, 2014

ISSN 1914-525X
Catalogue No. M41-10/7-2014E-PDF
ISBN 978-1-100-22976-8
doi:10.4095/293327

A copy of this publication is also available for reference in depository libraries across Canada through access to the Depository Services Program's Web site at http://dsp-psd.pwgsc.gc.ca

This publication is available for free download through GEOSCAN
http://geoscan.ess.nrcan.gc.ca

Recommended citation

Dhesi, P., Haggart, J.W., and Brent, T., 2014. Effective project-level information management: development of a standardized protocol for capturing metadata within the Baffin Bay Basins Project of the GEM-Energy Program; Geological Survey of Canada, Technical Note 7, 7 p. doi:10.4095/293327

Critical review
A. Moore

Authors

P. Dhesi (Parm.Dhesi@NRCan-RNCan.gc.ca)
J.W. Haggart (Jim.Haggart@NRCan-RNCan.gc.ca)
Geological Survey of Canada
1500-605 Robson Street
Vancouver, British Columbia
V6B 5J3

T. Brent (Tom.Brent@NRCan-RNCan.gc.ca)
Geological Survey of Canada
3303 33rd Street NW
Calgary, Alberta
T2L 2A7

Correction date:

All requests for permission to reproduce this work, in whole or in part, for purposes of commercial use, resale, or redistribution shall be addressed to: Earth Sciences Sector Copyright Information Officer, Room 622C, 615 Booth Street, Ottawa, Ontario K1A 0E9.
E-mail: ESSCopyright@NRCan.gc.ca

Effective project-level information management: development of a standardized protocol for capturing metadata within the Baffin Bay Basins Project of the GEM-Energy Program

P. Dhesi, J.W. Haggart, and T. Brent

Dhesi, P., Haggart, J.W., and Brent, T., 2014. Effective project-level information management: development of a standardized protocol for capturing metadata within the Baffin Bay Basins Project of the GEM-Energy Program; Geological Survey of Canada, Technical Note 7, 7 p. doi:10.4095/293327

Abstract: Although often neglected, the generation of metadata is a critical component of any project information management process. Ideally, a corporate standard should define the type of metadata content to be collected for all project-related geospatial files. In practice, metadata collection is typically ad hoc and up to individual practice and preference. The authors present a simple method by which metadata can be easily initiated by the project participant in the course of generating geospatial data files. The original data files with the associated metadata files are then transferred to the project information management leader for vetting and long-term archiving. Perception of metadata generation, and its general acceptance of value by working-level scientists, can benefit from simply defined practices that can be easily executed. The authors' methodology is a relatively simple and streamlined implementation process that meets the authors' information management duty to create metadata, while exploring practical and understandable procedures, and is reflected in a real, project-level case study. The methodology and procedures can stand for consideration within the broader long-term determination of best practices or development of standards for information management within Natural Resources Canada.

Résumé : Bien qu'elle soit souvent négligée, la production de métadonnées est un élément crucial de tout processus de gestion de l'information. Dans un monde idéal, une norme organisationnelle devrait définir le type de métadonnées qui seraient assemblées, pour tous les fichiers géospatiaux liés à un projet. En pratique, la collecte des métadonnées varie avec les circonstances, ainsi que les pratiques et les préférences individuelles. Nous présentons une méthode simple grâce à laquelle le participant à un projet peut facilement entamer le processus au moment même où il produit des fichiers de données géospatiales. Ensuite, les fichiers de données originaux et leurs fichiers de métadonnées associés sont transférés au chef de la gestion de l'information du projet aux fins de vérification et pour l'archivage à long terme. La perception de la production de métadonnées et l'acceptation générale de leur valeur par les scientifiques directement impliqués dans la réalisation des travaux pourra profiter de pratiques définies simplement et d'exécution aisée. Notre méthodologie consiste en un processus de mise en œuvre relativement simple et rationalisé qui respecte notre devoir de gestion de l'information lié à la création métadonnées, tout en explorant des procédures pratiques et compréhensibles, et qui est illustré par une étude de cas réelle au niveau d'un projet. On devrait considérer notre méthodologie et nos procédures lorsque l'on déterminera dans une perspective plus large à long terme des pratiques exemplaires ou lorsque l'on élaborera des normes pour la gestion de l'information à Ressources naturelles Canada.

INTRODUCTION

By their nature, Earth science data are inherently geographic, three-dimensional (four-dimensional, if the time element is considered), and thematically diverse and such data are synthesized and interpreted spatially in the modern workflows of industry and government scientists. Metadata are critical to serve the many practical and scientific tasks for which they are commonly used. These include enhancing data browsing and discovery, effective data transfer and reformatting and/or editing and compilation, and data selection and referencing, to name a few. Metadata can be organized into classifications ranging from a simple listing of basic information about geospatial data to detailed documentation about an individual data set. At a fundamental level, metadata support the creation of an inventory (public and/or internal) of any data set. Metadata are also important in the creation of a spatial data clearinghouse, such as Earth Science Sector (ESS)'s GEOSCAN or GeoGratis interfaces, where potential users can search to find the data they require for their intended application. Metadata ensure that potential data users can make an informed decision about whether data are appropriate for the intended use. Metadata also help to ensure that the spatial data holdings of ESS are well documented and that the risk of data and knowledge loss is minimized when system and software upgrades are implemented, or when trained staff are reassigned.

Identification and recording of metadata present challenges, however, not the least of which is a lack of standards implementation. As noted by A. Yeung (Earth Sciences Sector Working Group on Archiving and Preserving Geospatial Data, 2005) for a closely related aspect of information management, the actual preservation of digital data, "...the real solution for digital preservation may lie less in technology and more in policy." Thus, an informed corporate policy applying to information management (IM) is first required, in order to encourage individuals or groups within the organization to support the defined objectives of information management and to develop corporate policy standards with respect to information management in the archiving of their own data sets.

CORPORATE INFORMATION MANAGEMENT PRIORITIES

The Geological Survey of Canada (GSC)'s priorities for dealing with information management are directed at three principle questions:

Who are our clients?

What do they want? and

How do they want it and how should we give it to them?

Individual projects within GSC often generate large and disparate data sets, for which answers to the above priority queries may not be intuitively available. It is thus critical that individuals generating such data files follow agreed-upon standards for data management, so that they can be suitably archived such that retrieval of the data for future use can be undertaken readily. In order to efficiently address these priorities, GSC is initiating activities to define and implement corporate standards for information management that meet Natural Resources Canada and Treasury Board of Canada requirements, and also create information management and metadata-generation workflow templates that can be used by project participants to enter metadata for future information applications, including archiving and retrieval. Additionally, a suitable suite of tools and procedures must be made available for metadata generation and archiving to specific programs.

At present, Treasury Board of Canada has adopted the ISO Standard on Geospatial Data that addresses critical aspects of metadata generation, collection, and archiving (*see* Appendix B: Implementation Conditions for ISO 19115 of the Treasury Board Secretariat Policies; http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=16553&section=text#); however, processes for easily completing metadata have not yet been implemented within the GSC. This paper describes a method of metadata accumulation at the project level that can ultimately be integrated within a corporate GSC platform for metadata capture and storage.

INFORMATION MANAGEMENT OBJECTIVES FOR THE BAFFIN BAY BASINS PROJECT

Participants within the Baffin Bay Basins Project of Earth Science Sector's Geoscience for Energy and Minerals (GEM)-Energy Program have generated a large number of geospatial and other digital data files during the execution of the project, including project presentations, administration, and data files. Organizing these files in an informed manner for long-term retention, as well as for shorter term analysis phases, has proved challenging, so the general objective has been to implement information management practices that facilitate efficient science-product generation from project data by the GSC, both now and in the future. This is critical with respect to the Baffin Bay Basins Project as industry and environmental interest in this region remains high and GSC science programs in the region have been renewed (i.e. GEM II).

As corporate procedures and mechanisms for information management within GSC have not yet been formally established and disseminated, the authors have developed an example framework for information files generated within the Baffin Bay Basins Project (GEM-Energy Program). This framework includes standard procedures and tool sets applicable at the project level to ensure that the Baffin Bay Basins Project outputs meet Treasury Board of Canada mandated

standards for data dissemination and metadata, and can serve as an example for other projects within the GEM Program, as well as other Earth Science Sector programs.

Objectives of information management and metadata generation initially identified for the Baffin Bay Basins Project included:

- establishing an internal project-specific corporate spatial data 'warehouse' that is managed as a read-only ESRI Spatial Database Engine (SDE) and/or Oracle database and only given write access with permission from the project information manager;
- implementing 'best practices' standards that have been defined for Spatial Database Engine layer naming, symbology, metadata requirements, and access security;
- identifying and transferring appropriate data assets to the spatial warehouse;
- ensuring data files are accessible to users via ESRI-based tools and/or other GIS desktop tools; and
- the creation of a metadata-generation template for project files that can be readily populated by nontechnical project participants.

METADATA STRUCTURING

Metadata for Baffin Bay Basins Project files has been organized in a structure that currently supports the Federal Geographic Data Committee (FGDC) standard and will be eventually migrated to the North American Profile (NAP) of ISO 19115:2003 – Geographic information – Metadata.

The North American Profile is comprised of eleven major metadata classes, each of which includes subitems and/or additional classes and attributes. A short description of each major class is provided in Table 1. There are 50 mandatory metadata elements in the ISO North American Profile.

The Canada Centre for Mapping and Earth Observation of the Earth Sciences Sector has developed a subset of the ISO NAP metadata. This subset has predefined values for more than half the elements, which can be selected from pick-lists or are generated automatically, thereby leaving only 18 mandatory metadata elements to populate with free text (and many of these are repetitive if multiple data files are being referenced). In practice, the most time-consuming metadata items to populate are the abstract, purpose, and supplemental information areas; however, these can be derived from other GSC processes such as the Publishing Process Integration (PPI) system and the new Departmental Science and Technology Publications Policy, which requires a plain-language summary of the NRCan Science and Technology publication as well as potential science and policy implications. This summary is also stored and accessed through the Publishing Process Integration system.

Table 1. Principal metadata classes and attributes addressed in Baffin Bay Basins Project.

Major class name	Description
Metadata	A description of the overall metadata record
Identification	Basic information about the data set
Metadata constraints	Describes the use of the metadata and legal and security constraints on this use
Data quality	Data quality information for the resource
Metadata maintenance	Information about metadata update
Spatial representation	Digital representation of vector and/or grid objects in the data set
Reference system	Description of the spatial and/or temporal reference systems used in the data set
Content	Characteristics describing the feature catalogue, the coverage, and the image data
Portrayal catalogue	Information about the catalogue describing symbols and rules to depict a resource
Distribution	Information about acquiring the data set
Application schema	Information about the conceptual schema of the data set

BASIC METADATA CREATION GUIDELINES

Metadata generation is typically straightforward if initiated at the early stages of the data-generation process. The longer the metadata-generation process is delayed, the more difficult it becomes to recall details about the data, thereby affecting its reuse. A suggested process to follow in the generation of metadata is as follows:

- write an informative title for the file, providing a subject or theme, place, NTS sheet, time, and purpose;
- write an informative abstract no longer than 250 words;
- document constraints on use of the file appropriately;
- determine access constraints, if applicable;
- assess use constraints, if any;
- identify effective keywords: think both broadly and narrowly so that others with different objectives can find the record (dictionaries such as the GCMD directory [http://gcmd.nasa.gov/learn/faqs/search_explanation.html] can help populate the keywords); and
- verify the metadata to ensure correct spelling and accuracy: various tools in the metadata-generation template help to standardize these aspects.

METADATA CAPTURE TOOLS

To encourage and promote efficient capture of metadata, the Baffin Bay Basins Project utilized specific software and tools. These include:

- ESRI Desktop Suite (ArcMap and ArcCatalog);
- Safe Feature Manipulation Engine (FME);
- EME (EPA Metadata Editor) toolset for ArcCatalog; and
- Microsoft Excel.

All metadata for geospatial files can be created in the ESRI product called ArcCatalog. Geospatial files and their metadata can be stored locally or be maintained in an ESRI Spatial Database Engine (SDE) that can allow others with appropriate permissions to access the data. ArcCatalog has a simple metadata editor built into the product, but it is not sufficiently robust and does not provide the semiautomation necessary to quickly and efficiently update and validate metadata.

To remedy this limitation, a tool created by the U.S. Environmental Protection Agency called EPA Metadata Editor (EME) has been incorporated into the metadata generation process for the Baffin Bay Basins Project. The EPA Metadata Editor is a smart three-tab editor toolset installed within ArcCatalog for creating, editing, and validating metadata. It is very user-friendly, colour-coded for defining field importance, and is database driven so that users can create their own predefined code values. A second tool utilized, called Feature Manipulation Engine (FME), provides a method to batch process many metadata attributes if they are common among many spatial data sets. A third option, for clients that do not have access to ArcCatalog or Feature Manipulation Engine or who prefer a simpler metadata generation process, utilizes dropdown menu selection choices within a customized Excel spreadsheet. Project participants can easily and readily capture the relevant metadata for their project files without the need for a third-party tool. Figure 1 represents screen captures of the basic data capture fields that are incorporated in EPA Metadata Editor, as modified for the Baffin Bay Basins Project. If clients are entering metadata through the Excel spreadsheet, principal areas of metadata captured include 'Basic Data Set Information' (Fig. 1a), 'Quality, Coord Sys, Att. Info'(Fig. 1b), and 'Distribution' (Fig. 1c).

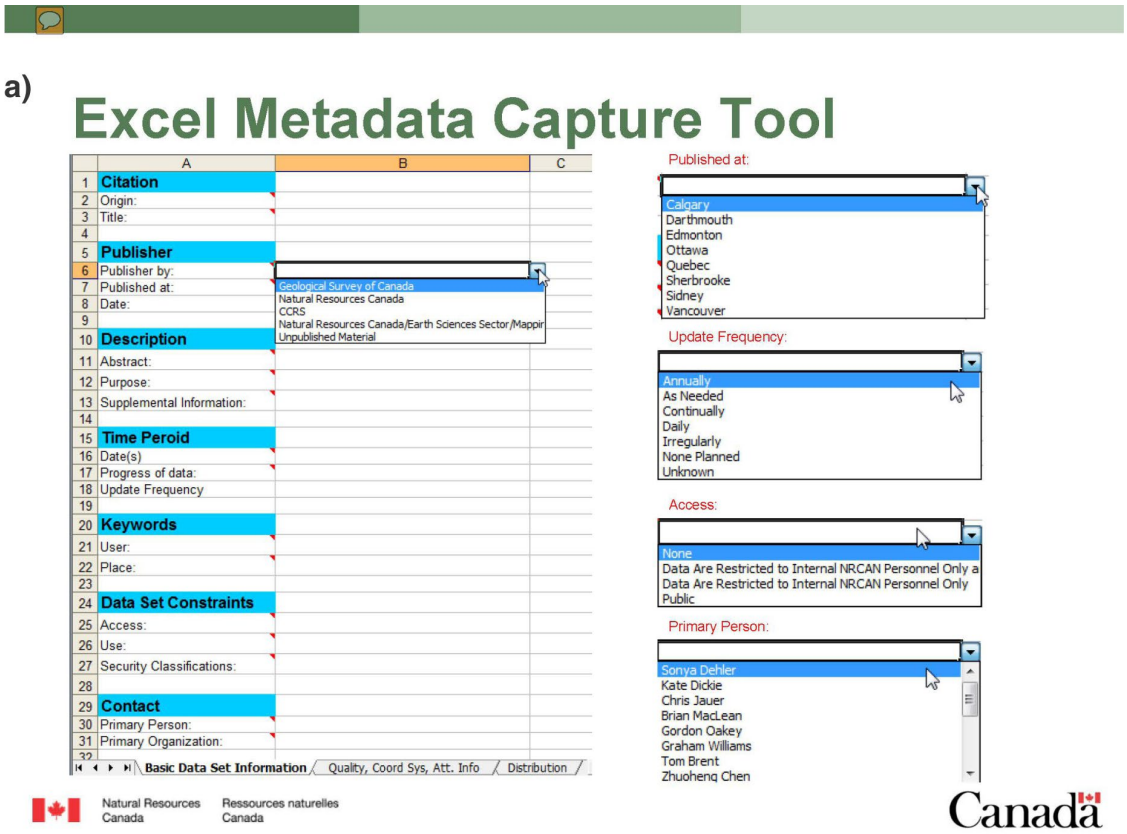

Figure 1. Three-tabbed metadata capture tool designed for Baffin Bay Basins Project that allows project participants to populate their spatial data files. The tool is designed to capture the most important aspects of metadata for the spatial data set. The spreadsheet is analogous to the metadata-capturing process in ArcCatalog, but in Excel format instead, which provides ease of entry for clients who do not have access to ArcGIS tools or are not fluent in their use. **a)** 'Basic Data Set Information' tab: screen display of fields captured within 'Basic Data Set Information' tab.

PROJECT METADATA VETTING

Once relevant metadata for geospatial Baffin Bay Basins Project files are generated by the project participant, the files are passed to the project information management leader for eventual archiving. The information management leader ensures that both the file and the associated metadata are suitable for long-term corporate archiving and maintenance, and that both utilize suitable quality control and validation. In addition, the information management leader assigns appropriate sharing permissions for the file and verifies that basic metadata have been generated, including titles, points of contact, licence constraints, keywords, attribute information, projection information (geospatial co-ordinate system), and dates. The responsibility for the overall quality of the metadata is shared between the information management leader and the scientist, as aspects of the abstract, purpose, and supplemental information greatly influence the ultimate value of larger metadata goals. It is understood that metadata are editable and that when additional insights or recognition of inaccurate or metadata mistakes are discovered, these can be accommodated. The vetting process is considered an iterative process in this implementation, so as to promote quality.

When metadata have been generated, the project participant uploads relevant files for archiving to a specific directory under the departmental intranet shared drive assigned to the participant. These files are shown conceptually within the yellow field of Figure 2, under 'Client Data'. In addition to geospatial files, other types of files can be included here, including spreadsheets, poster and presentation PDFs, PowerPoint files, and project administration documents. The participant then notifies the project information management leader that the files have been placed within the specific client data directory.

The project information management leader then reviews the data files and the associated metadata, checking that all required data fields have been captured and that the data files

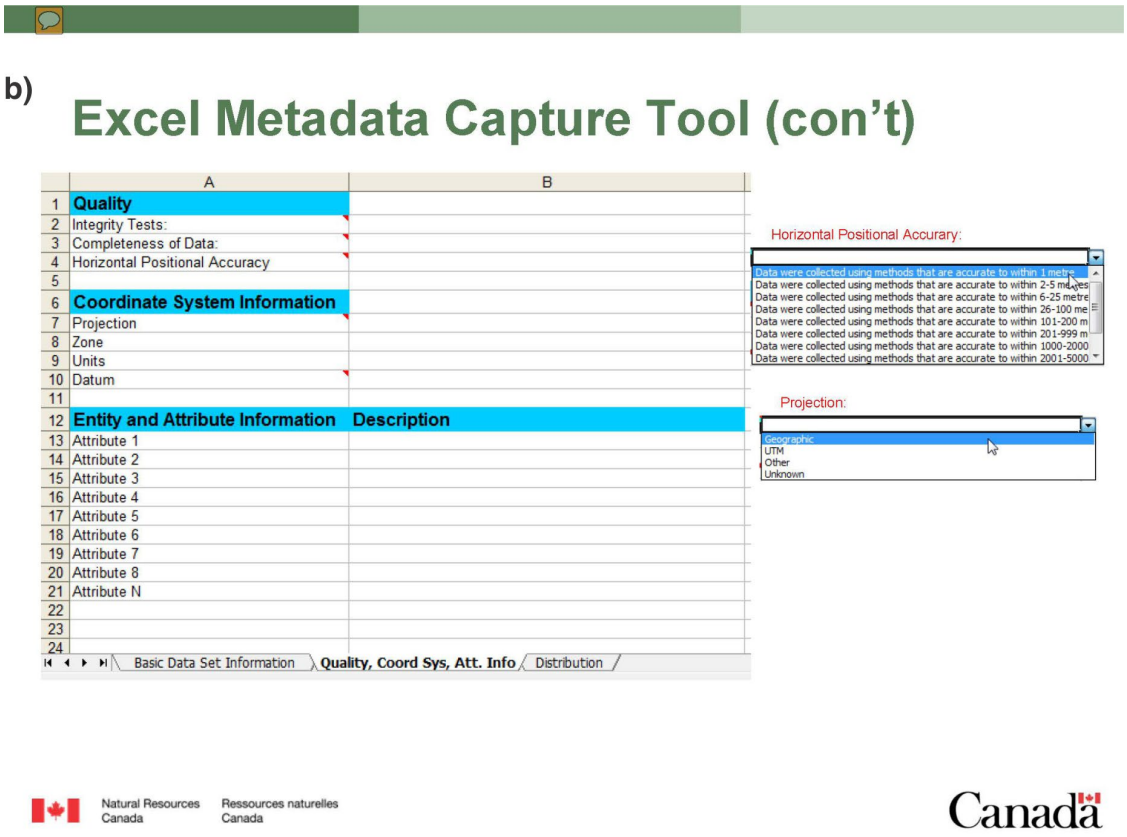

Figure 1. (continued) **b)** 'Quality, Coord Sys, Att. Info' tab: screen display of fields captured within 'Quality, Co-ordinate System, Attribute Information' tab. Under 'Entity and Attribute Information', users can enter any field names and provide a brief description of the field. Hovering the cursor over the small red triangle in each field gives a brief description of the kind of information being requested.

themselves are properly configured for long-term corporate archiving. Once outstanding issues have been addressed and the project leader has given his/her approval, the files are then transferred to the corporate directory for permanent archiving.

OUTSTANDING ISSUES AND CONCLUSIONS

This contribution describes efforts to organize data generated by the Baffin Bay Basins Project of the GEM-Energy Program. It has been developed at the project level to deal with the disparate and large data sets generated from the project. Discussions with Earth Science Sector corporate information management groups have provided guidance for the development of the methodology, but in the absence of clearly defined corporate goals, guidelines, and tools, the authors' methodology must be considered preliminary in nature. Nonetheless, the authors believe it provides an effective information management tool that enhances the value of data files at the corporate level.

At the time of this writing, several issues and concerns remain that prevent the enhanced utilization of this metadata-capture methodology.

Science projects within the GSC have not received directives on information management, nor have they been provided with the necessary tools to organize their spatial data assets and other information. With organizational changes in Earth Science Sector over the past number of years, it has only now become clear that the GSC has the responsibility to define its information management infrastructure, including how to implement Government of Canada metadata standards. The development and implementation of metadata procedures by information management personnel and researchers, in order to appropriately manage their project information data, has been regional in nature and unco-ordinated across the GSC. Such an approach inevitably results in multiple methodologies that are difficult to integrate at the corporate level. A co-ordinated and common approach is thus required across the GSC to record project information, namely spatial metadata.

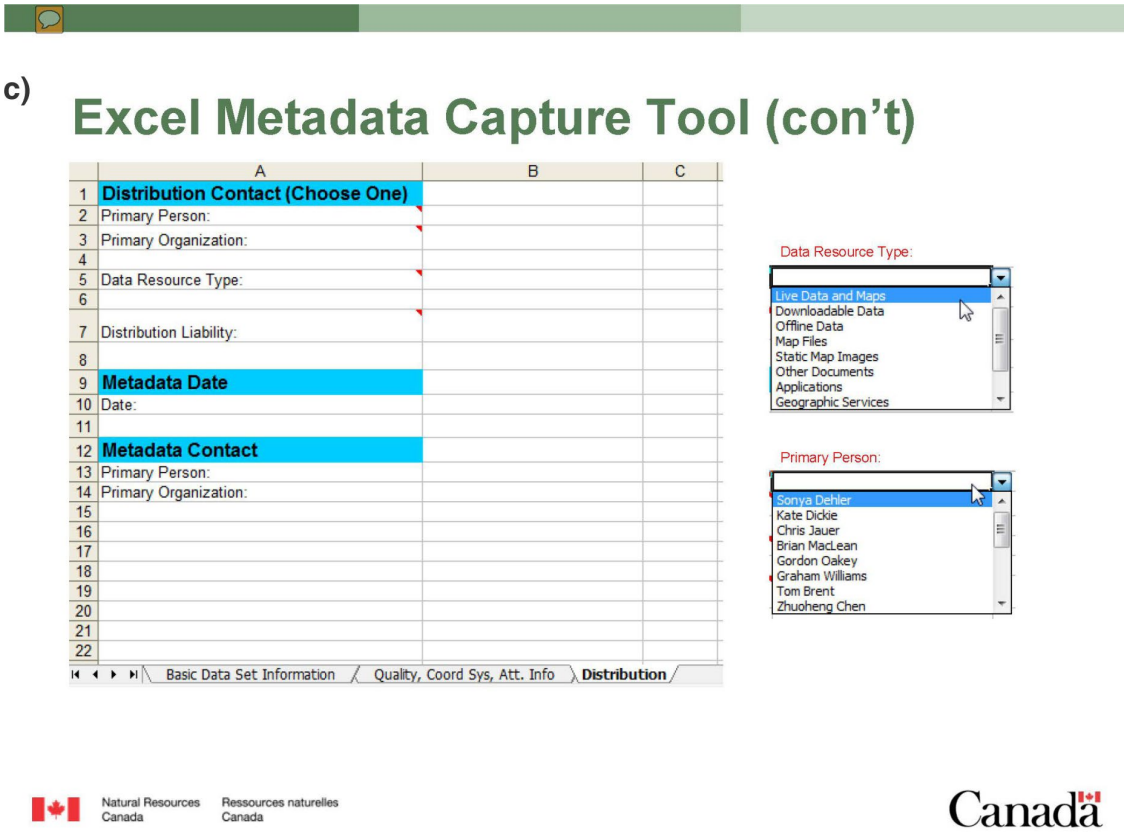

Figure 1. continued) **c)** 'Distribution tab': screen display of fields captured within 'Distribution' tab. This tab captures all primary contact information for the file. Participants' names have been added so users just need to select their names from the dropdown menu. The editor is modified from U.S. Environmental Protection Agency software EPA Metadata Editor.

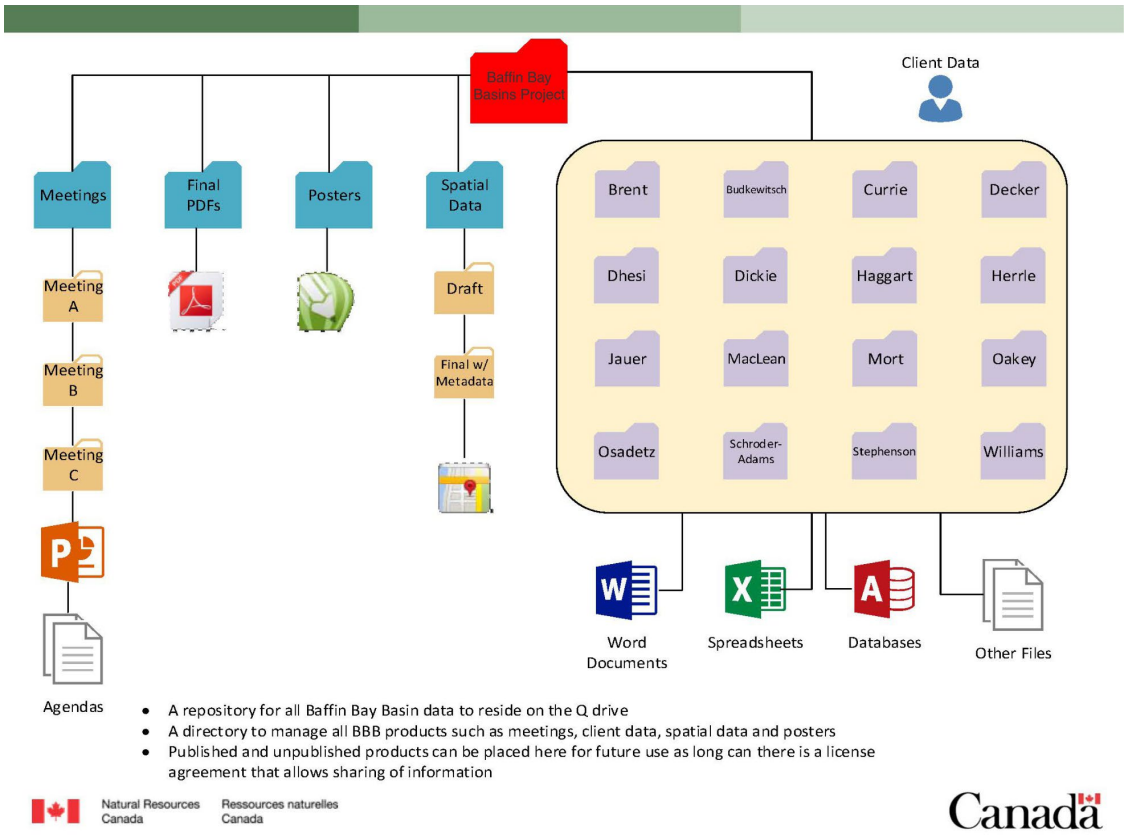

Figure 2. Baffin Bay Basins Project repository. A structural representation for all Baffin Bay Basins Project data to reside on corporate storage devices. A system to manage all Baffin Bay Basins Project products such as spatial data, meetings, client data, and posters. Published and unpublished products can be placed here for future use, provided there is a licence agreement that allows sharing of information.

Each project requires a keyword list to populate metadata attributes for any given layer. Such lists will provide consistency between common spatial data types and speed up the production of metadata capture. Initially, such keyword lists will be project- or subject-specific, but over time, it is likely that many of these keyword lists can be standardized at the corporate level for use by most, if not all, projects.

Metadata entry is typically left to the end of the spatial generating process. This can have significant repercussions as information may be lost over the life of the project and the human resources for collecting and generating the data may be lost through staff attrition and transfer. Thus, it is imperative that metadata capture take place very early in the data generation process and continues and is updated throughout the life of the project.

The authors' approach to metadata generation and management can be considered as an example of how metadata can be collected and archived for subsequent information management use. The authors suggest that such procedures should be considered for broader implementation at the corporate level and, in fact, some effort to do this has been initiated within the ESS-GSC information management group. Efforts have been made by the authors to communicate with that group and others to share information and ideas to ensure that efforts are aligned with the broader corporate framework.

ACKNOWLEDGMENTS

The authors thank project participants of the Baffin Bay Basins Project of Earth Science Sector's GEM-Energy Program for informed discussion about metadata issues and collection challenges. P. Moir provided particularly stimulating discussion about information management philosophy early in the project process. A. Moore and A. Laviolette are thanked for succinct comments on the manuscript and for providing a higher level corporate perspective on metadata management.

REFERENCE

Earth Sciences Sector Working Group on Archiving and Preserving Geospatial Data, 2005. Archiving, management and preservation of geospatial data. Summary report and recommendations. Canadian Geospatial Data Infrastructure Information Product 23e, <http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/38389f5e-19f3-5093-9742-4ccecae16039.html>

Geological Survey of Canada Project EGM005