Mapping Out a Region of Interacting Standards

Thomas, Wendy

University of Minnesota, Minnesota Population Center

50 Willey Hall

225 19th Avenue South

Minneapolis, MN 55455, USA

E-mail: wlt@umn.edu

Introduction

Official statistics cover a wide range of topics including demographic description, social, political and economic activity, and ecological and environmental. The this information is used in analyzing change and interaction to inform decision making and policy development as well as related research. In order to analyze these data, metadata must be accessed, exchanged, and processed by many different forms of software and systems. Standards for metadata and data storage and transfer facilitate this process. A number of standards have been developed to address very specific basic features such as the description of a data element (ISO/IEC 11179) structured to facilitate registration and management of data elements over time in a way that they can be shared within and between organizations. Others address discovery, collection management, and exchange. For example, Dublin Core provides a high level generic bibliographic record to support searching and collection management. The Metadata Encoding and Transmission Standard (METS) focuses on exchange by providing a wrapper which informs the recipient of the structure and content of the packaged material. Broad models like the Generic Statistical Business Process Model (GSBPM) focus on the process flow of procedures and common steps in the creation of statistics within statistical agencies.

Detailed metadata content models generally focus on areas of data collection and specialization. They tend to have considerable content overlap but each focuses in on the special informational issues of their topic area. As evidenced by recent IAOS conferences such as the "IAOS Conference on Official Statistics and the Environment: approaches, issues, challenges and linkages", in Santiago, Chile October 2010, there is increasing demand for linking data from multiple topical areas as well as data from a variety of sources (survey, administrative, statistical, qualitative, monitoring systems, etc.). Relating data from the traditional areas of official statistics (social, economic, labor, health and financial data) to areas of environmental, climate, oceanographic, land use and related data or to other specialized data sources becomes an issue of effectively moving between multiple standards without loss of critical content.

Many of these areas have developed standards for capturing metadata in structures that support analysis within their domains. As expanded analysis crosses these domain lines, smoothly interacting metadata standards will improve understanding and accurate use of data from related domains. Understanding the critical points of interaction between major standards will be the key to improving the analysis of data in the public sector. The purpose of this paper is to map out a region within the area of official metadata and statistics where standards need to interact to support the process of data design, capture, analysis, and dissemination within and between related domains.

Key areas of interaction

The value of standards is known. Standards support exchange of data between independently designed systems and software. Some of these systems interact with selective objects within the standard, mapping them to their own internal structures and importing them for processing. These systems may also export the limited set of objects they use, however, additional metadata from the imported metadata is lost or new metadata created by the system cannot be reincorporated into the original metadata. These systems may be sufficient for accessing and delivering data and metadata to users but is insufficient for managing the metadata throughout its lifetime in a consistent and effective manner. There is no one standard that manages all aspects of data: conceptual design, process management, collection, testing, evaluation, data analysis, dissemination, preservation, and integration with other data types. Different standards support different aspects or stages within the lifecycle of the data. In addition, different topical and application areas have created different standards providing extended details where needed. The goal is to identify the key areas of interaction between standards so that as data and metadata move through various stages or interact with data from other disciplines, the user can create clear and accurate links, process the metadata within a specific discipline area or standard without information loss, and then integrate new data and metadata that is created during the processing of the original materials.

No two standards map one-to-one on all of their content. However, many standards are designed with the intent to interact effectively with other standards in terms of their technical design and coverage overlap. For example, the Data Documentation Initiative (DDI) looked extensively as over 30 standards in developing its DDI Lifecycle model (DDI 3 and beyond). It included an ISO/IEC 11179 based model for describing data elements, incorporated native Dublin Core elements as well as a more detailed citation structure, included structural information needed by geographic systems to link data to specific geographic features using ISO 19115 (Geographic Information) as a model, and modified its aggregated data structures to map cleanly to SDMX for use as an output structure or as a clean exchange to this standard for continued data/metadata management. DDI Lifecycle model was used as one of the inputs to the GSBPM structure and there are strong similarities as the top levels of these structures.

The assumption is that during its lifetime, metadata will pass through many different processes, types of software, and standards. Although most production systems follow similar general steps the specific internal processes, the software used to management and process data/metadata, and the standards used to capture the metadata varies over the entire process and between organizations. In determining the suite of standards to use throughout the life of data, a number of points must be considered.

- What is required for consistent identification:
    - Persistent identifiers for study, data files, data elements and metadata objects
    - Can these identifiers be transferred between systems/standards in a way that prevents loss
- What metadata is required to be in a specific structure to drive software (machine-actionable metadata)
- What is required at different stages of the data/metadata lifecycle:
    - Process management
    - Use of materials from outside of specific data collection (concepts, definitions, etc)
    - During the production process
    - Post-publication and/or distribution (thinking about how the material will be used)
- What metadata is generated by specific software
- What information is needed to link collected data to other data sources

Technical points of interaction

Identifiers have different purposes in different systems. Ideally an identifier should be persistent and unique. ISO/IEC 11179 uses a full identifier composed of three parts: agency, identifier, and version. This allows for management of identifiers within an agency such as a national statistical agency and capturing version changes of an object over time. Some systems use the services of agencies providing unique identifiers and indexing of the object's location. They may or may not incorporate versioning. The point being that different standards and different systems do not all use the same model for identification. When metadata content must move into and out of different systems, software, and standards models, the variation in identification systems must be accounted for to prevent loss of information as metadata passes through an alternate structure. At minimum, there should be a means of storing the non-conforming identification information is a format that retains its link to the original object while allowing the standard to supply a local identifier for use within that specific environment.

Metadata objects are generally either machine-actionable (tightly structured in order to drive software) or informational (designed to be presented and read by the user). Machine-actionable metadata may be defined by the standard or be a system or software imposed constraint on the content. Metadata may be compliant with a standard and yet not meet additional constraints of the software. Standards that support a structured means for adding system specific constraints assist in clarifying requirements for moving metadata into and out of different systems. DDI provides an example of two means of capturing these additional constraints. First is the provision for the use of external controlled vocabularies and second, the DDI Profile which lists the coverage and constraints of an applications use of DDI.

Coverage interaction

A process model such as GSBPM provides a base for mapping the interaction points and potential coverage gaps in various standards. Extending this model to look at the potential linkages of the collected data to other data sources as well as the long term preservation needs within an archival organization allows for mapping metadata capture standards such as ISO/IEC 11179, DDI, SDMX, ISO 19115, and Dublin Core against various stages in the data/metadata lifecycle and then determine potential points of metadata transfer between standards based on processing and usage. For example, one standard may capture process steps and track the movement of a data collection process from needs specification through process evaluation. This metadata collection may parallel the collection of metadata that is more focused on the resultant data, describing overall processes but not tracking and managing the process itself.

At certain points in the process metadata may be integrated from systems designed to capture and manage common structures such as concepts, universe descriptions, and data elements. These may be held in an ISO/IEC 11179 compliant registry whose management and update is driven by a sub-process in the overall metadata process management system. Finalized outputs may be both DDI based and SDMX based depending on the content of the data and needs of the user. In addition, the process may require the use of a GIS system, integrating the data with spatial files for analysis of the data or the output of certain products. Ensuring that the metadata required for transparent linkages between the data and its spatial designation is present provides support both for planned products and future uses of the data.

As part of the overall process data and associated metadata may be subject to a disposition requirement that deposits it in a separate archive where additional metadata may be added directly to the deposited metadata within the standard structure in use, or bundled within a METS-like object as a parallel set of information. Anticipating future metadata needs eases the movement of metadata between systems and reduces the possibility of metadata loss.

The analysis of standards interaction across a process also serves as a base for examining the interaction of software with metadata creation. While some software may merely process through metadata, accepting and processing commands and outputting data, other software creates metadata that needs to be reintegrated into the metadata store. Ideally the system should ensure that metadata is

captured at the source be that the entering of program into software or the interactive processing of data while within the software environment.

Implications for the content and coverage of the data we collect

Standards can facilitate the exchange of data within and between various applications and a range of disciplines. However, standards by themselves do not insure the accurate and efficient use of data within different areas of research, geographic detail vs. confidentiality, point vs. polygon data, and national policy vs. local programs. In taking the broader view of potential data usage consideration can be given to new types of data products that can be generated by systems driven by metadata. For example by capturing processes for aggregation and the application of confidentiality constraints, products can be generated on-demand (within access constraints).

The intent of process management and the use of standards to create metadata driven processes is to both increase efficiency and improve consistency. It also raises the possibility of improved access and more targeted data products. The ability to accurately and easily assess the temporal, spatial and topical links between data sets allows deeper inquiry into the interactions between a range of disciplines particularly in the discipline groups of social, behavioral, and economic studies and ecological, land use, and oceanic studies. Standard techniques used in the first group to protect confidentiality may need to be reconsidered in order to meet the more spatially specific needs of the second group. Increasing the options for spatial interpolation by providing a limited set of variables for small sub-areas within areas with commonly published aggregate data could improve research exploring the interactions between these discipline areas.

Metadata standards can be used not only to capture, preserve and deliver metadata to researchers, but can be used to better plan and manage the data creation process and to anticipate where other researchers may require specific metadata to meet their needs. By examining the lifecycle of data and metadata as a whole, we can anticipate need, collect useful metadata at the source and improve both access to data and reliability of research results.

REFERENCES (RÉFERENCES)

Bargmeyer, Bruce E., Daniel W. Gillman, "Metadata Standards and Metadata Registries: and Overview", Bureau of Labor Statistics, Washington, DC 20212 http://www.bls.gov/ore/pdf/st000010.pdf

Gregory, Arofan, Pascal Heus."DDI and SDMX: Complementary, Not Competing, Standards", Open Data Foundation Paper, July 2007. http://odaf.org/papers/DDI_and_SDMX.pdf

Gregory, Arofan, Pascal Heus, Jostein Ryssevik, "Metadata", Council for Social and Economic Data (RatSWD), Working Paper 57, March 2009 http://www.ratswd.de/download/workingpapers2009

Vardigan, Mary, Pascal Heus, Wendy Thomas. "Data Documentation Initiative: Toward a Standard for the Social Sciences." The International Journal of Digital Curation 3, 1 (2008).

RÉSUMÉ (ABSTRACT)

Discussions of metadata standards for organizing, exchanging and preserving metadata no longer focus on the value of supporting such standards or which one standard should be used. The focus is now on how standards interact, supporting varieties of search systems as well as the movement of metadata and data through different standards to support specialized use of the content. Discussion now focuses on a suite of standards addressing different types of data at different points in their lifecycles, being captured and processed for different purposes. Data collection processes described in the Generic Statistical Business Process Model (GSBPM), may capture metadata using the Data Documentation Initiative (DDI), process their microdata into statistical tables and distribute it using the Statistical Data and Metadata Exchange (SDMX). As this output moves into digital libraries it may be placed inside a Metadata Encoding and Transmission Standard (METS) wrapper for exchange within that system, searched via a Dublin Core browser, or have preservation metadata added to it. Others working in a

geospatial environment may then selectively transport data and metadata into related geospatial metadata (ISO 19115 based).

The purpose of this paper is to map out a region within the area of official metadata and statistics where standards need to interact to support the process of data design, capture, analysis, and dissemination within and between related domains. The "IAOS Conference on Official Statistics and the Environment: approaches, issues, challenges and linkages", in Santiago, Chile October 2010, highlighted the need to relate data from the traditional areas of official statistics (social, economic, labor, health and financial data) to areas of environmental, climate, oceanographic, land use and related data. Many of these areas have developed standards for capturing metadata in structures that support analysis within their domains. As expanded analysis crosses these domain lines, smoothly interacting metadata standards will improve understanding and accurate use of data from related domains. Understanding the critical points of interaction between major standards will be the key to improving the analysis of data in the public sector.