

Toward Interoperability: A Report from the 11th Open Forum on Metadata Registries and Related Standards

by Gail Hodge

Information Standards

Today's standards are critical to a modern society. They support safety, ensure health and a clean environment, test our food, promote worldwide communications and trade and support innovation and economic development. Today's information management requires attention to a vast array of standards ranging from network protocols to format and content standards. Where once information technology standards were the purview of programmers and system administrators, and librarians and information specialists were only concerned about content standards, the current environment demands that information managers turn their attention to all types of standards.

A standard is an agreed upon way of doing something. Standards are of several types and serve multiple functions. Standards may be applied to physical items or units of measure; terms, definitions, classes, grades, rating or symbols; test methods or recommended practices; systems and services, in particular quality and related aspects of management system standards; and standards for health, safety, consumers and the environment. Standards are international in scope, such as ISO, or regional in scope, such as those of the EU. They may be voluntary or required by a particular industry, government or international treaty. Standards may begin as voluntary or ad hoc standards and then develop into formal standards as they are adopted. Standards may be agreed upon by a very small group or recognized worldwide. Standards occur in a multitude of disciplines and enterprises from defense to manufacturing to agriculture and the social sciences.

However, despite the importance of standards in information

Gail Hodge is senior information scientist with Information International Associates, Inc. She can be reached by mail at 312 Walnut Place, Havertown, PA 19083; or by email at ghodge@iiaweb.com

management, they can lead to confusion rather than clarity. Information managers are often overwhelmed by the number of standards. Few information systems can be specified, procured, developed or implemented without using multiple standards. Some standards appear to compete and in other cases there are gaps between what the standards cover and the requirements of the information system. How do all these standards fit together? Which standards should I use for which purpose? While many conferences and workshops tend to focus on a single standard or family of standards, this approach may not present the breadth of scope that practitioners need. On the other hand, agenda that focus solely on implementation of standards in a particular domain such as healthcare or manufacturing, do not give participants the benefit of learning "outside their box."

Background on the Open Forum on Metadata Registries

With the goal of bridging various standards and practices in the information management domain, the 11th Open Forum on Metadata Registries was held on May 19-22, 2008, in Sydney Australia. "Metadata DownUnder: Metadata, Semantics and Interoperability" focused on metadata, registries, semantics and related standards and information management practices (www.metadataopenforum.org). More than 100 practitioners, standards developers, information system developers, technologists, information managers and policy makers from more than 10 countries attended. The tutorials, track sessions, poster sessions and keynote presentations brought together people representing different standards, perspectives and application areas to share developments and exchange viewpoints.

The first Open Forum was held in Berkeley, California, in 1997. While early conferences focused almost exclusively on the ISO/IEC 11179 metadata registry standard, the conference has gradually expanded in scope. The

topics covered by the forum have changed as various standards have matured and the needs of the communities have changed to include practical implementation as well as standards development.

This forum provided information exchange across a wide range of information management standards and metadata application areas, covering metadata standards in use, current standards developments, new standards research and directions to support emerging needs. It covered a wide range of standards including geospatial, metadata, terminology, registry and data exchange. Application areas included transport systems, health, energy and the environment, manufacturing, statistics, government, defense and e-commerce. This forum described initiatives for bringing various standards groups together for the purpose of making standards more interoperable and, therefore, more useful and usable to information managers. As the promotional material suggested, this meeting was truly a “bazaar.”

This paper presents an overview of the meeting with a focus on the theme of interoperability. The slides for in-depth tutorials on a number of the standards are available from the Open Forum website. The last names of presenters are provided as appropriate. All presentations are available from the Open Forum website by clicking on “Past Forum” and selecting the 2008 meeting.

Interoperability Defined

Interoperability is a condition under which dissimilar entities, systems or, in this case, standards, can be interfaced. The goal is to connect the standards with a minimum of loss at the points of interface. In the case of standards, we are often talking about the need to harmonize the terminology and key concepts from the standard, interact with the different data models for each standard, and reconcile approaches and best practices.

In the context of this year’s Open Forum, interoperability is a goal at several levels. Standards themselves are developed, promoted and implemented to promote interoperability within and among systems, within an enterprise or across communities. In addition, there is increased attention being paid to improving the interoperability among the various standards.

Interoperability Drivers

Several key factors drive the need for interoperability of standards. These drivers include the vast amount of digital information, the increased diversity of content in digital form, the drive toward services, increased collaboration in all domains and user expectations.

The vast amount of information requires attention to standards of all types. It isn’t enough to rely on standards in one area, such as network communications, while ignoring the need to adequately organize and access the information itself through metadata and content standards. Putting more information in digital form only aggravates the current situation without improving it. The vast amount of information becomes more manageable and usable when other standards, such as standard metatags, controlled vocabularies for the content of appropriate metadata elements or a community-based taxonomy or a mark-up language, are brought to bear.

The increased diversity of content in digital form, from video to audio to virtual worlds, adds another dimension of complexity. When standards related to different content and format types become interoperable, the stage is set for single applications such as search engines to provide access to information regardless of the type or format.

The drive toward services and service-oriented architectures (SOA) is another factor driving the standards world to seek increased interoperability. While humans can cope with an amazing amount of ambiguity, computers cannot. Architectures based on standards allow computers to take on increased responsibility for the organization, management and even the analysis and use of information. This trend can be seen in the increase in computational science and other enhanced uses of computers. The use of standards to support SOA will increase as we move toward cloud computing, where the majority of the computation takes place on the network rather than on local devices. This technology allows devices to become even smaller in scale and integrated into other very complex but ubiquitous systems and products such as homes, cars and even clothing.

Collaboration is also driving the need for interoperability. Collaborators even representing different enterprises within the same domain must rely on

common standards and practices to work together smoothly and with a minimum of misunderstanding and error. Multi-disciplinary projects that occur in academic environments in the hard sciences, social sciences and the humanities and increasingly in our global economy require even more robust attention to methods for smoothing the path. If standards can be agreed upon, the teams can focus on the real goal of finding the common threads that will support the advancement of mutual goals.

Last but not least, a major driver for increased standards interoperability is that users are expecting – and even demanding – it. As stovepipes are broken down among enterprises, applications and groups, users do not want to concern themselves with the details of how to move from one standards platform or framework to another. Just as users are increasingly demanding open architectures for their information technology and applications, they are demanding openness in the standards that sit behind these technologies and applications.

Approaches to Achieving Interoperability

Many standards of importance to information management have been evolving over decades and are strongly entrenched in their individual communities. How do we achieve and implement interoperability among standards? The Open Forum presentations suggested several ways this can be achieved.

Community metadata profiles. Dublin Core is deeply embedded as a standard for librarians and information managers who deal with web resources. The Dublin Core was cited as the basis for resource description and discovery in existing systems and services described during the Open Forum [for example, presentations by Hunter; Hodge & Hutchison; Bromage]. At its most basic, the Dublin Core remains a very simple bibliographic standard for describing web resources. However, almost from its inception, the Dublin Core was found to be useful in ways its originators did not intend. Over the years, it has evolved to service a wide range of applications.

As the Dublin Core was applied to different types of resources in different subject domains and communities, such as archives and education,

it became apparent that more structure was needed to ensure more organized approaches to these variant applications [Sugimoto]. Therefore, a current focus of the Dublin Core Metadata Initiative is the development of Dublin Core Application Profiles. These profiles add, delete or qualify Dublin Core terms to address the needs of a particular content type or community and are intended to be based on community agreement. This approach allows for content discovery across domains using the core elements, while enhancing each community's ability to adequately describe resources to meet specific community needs. **[Editor's note:** Also please see the article on this topic by Nagamori and Sugimoto in this special section].

Metadata registries. Once the metadata schema has been agreed upon by a community, it is important to promote its use by announcing it as a best practice, norm or standard for those undertaking similar activities. It is also important to ensure that potential users can find the metadata schema and recognize it as authoritative. These functions can be served by a metadata registry. ISO/IEC 11179 has often served as the foundation for such registries.

ISO 11179 – Metadata Registry Standard supports the data community by registering data about data elements. The standard identifies metadata elements and registration processes to support the control, access and re-use of data elements within an enterprise [Fitzwater]. ISO 11179 has been used by organizations such as the European Environment Agency, the US Environmental Protection Agency, the Federal Aviation Administration and the National Cancer Institute (NCI). In the latter case, researchers or clinical study developers register their data elements in the caDSR and provide metadata about each data element [Warzel; Reeves]. When a developer or researcher begins another project, he or she accesses the registry and identifies data elements in the registry that are conceptually the same as those needed for the new study. Similarly, when collaborations are formed, the collaborators can use the registry to lay the groundwork for common understandings of what their joint research will entail. This is particularly valuable in external collaborations. Similar cancer registries have been created in Iceland and in the United Kingdom [Warzel; Harris]. In addition to shortening the development process, the re-use of accepted data

elements, their definitions and value domains supports NCI's ability to integrate data across studies and over time.

ISO 11179 has served as the basis for the development of registry components that support the development and use of other standards. For example, in an effort to organize and promote the re-use of the profiles that have already been developed, the Dublin Core community has developed a Dublin Core (DC) Registry [Sugimoto]. The DC Registry is loosely based on the 11179 registry standard, but it is much lighter weight in terms of the detail of the metadata required. The DC Registry includes the basic Dublin Core elements and other elements such as education level and biological organism that have been created by particular communities. The DC Registry can be searched and browsed from the DCMI website. APIs and web services are available so that other applications can display the metadata about each term. In addition to supporting profiles, the DC Registry supports interoperability across languages, storing 25 foreign language translations for the English DC terms. The DC Registry is distributed with sites in Japan, Germany, China and New Zealand.

The National Library of New Zealand Te Puna Mātauranga has implemented a DC Registry with Māori equivalents [Rollitt]. With the resurgence in the use of Te Reo Māori in New Zealand and its adoption as a second official language for New Zealand, the number of publications and resources in the Māori language increased. Having these terms well documented encourages more consistent use of the authoritative version of the DC terms. The library is planning to add other Pacific Island languages spoken in New Zealand and to register the Te Reo Māori Dublin Core as a New Zealand Standard. As a New Zealand Standard it will become a key component of New Zealand's Digital Strategy, a five-year plan to create a digital future for all New Zealanders with the goal of making New Zealand a world leader in the use of information and technology while continuing to support international standards and interoperability.

Shared vocabularies and knowledge organization systems. Shared vocabularies and knowledge organization systems are a key component of a metadata strategy, serving as the basis for many content standards and guidelines for completing the metadata elements. For example, the Dublin

Core, with its roots in the library community and standards such as MARC and AACR2, includes a subject field, which may have specific domain values taken from a controlled vocabulary or concept system, such as a thesaurus, classification scheme or subject heading list. However, appropriate controlled vocabularies and knowledge organization systems are often difficult to find, forcing developers to reinvent rather than reuse these resources.

To alleviate this problem the Food and Agriculture Organization (FAO) of the United Nations developed metadata for a registry of key concept or knowledge organization systems (KOS) as part of the Agricultural Information Management Standards (AIMS) website [Salokhe]. The objective of AIMS is to harmonize the decentralized efforts that are currently taking place in the development of methods and standards for information management and to facilitate collaboration and partnerships by promoting information exchange and knowledge sharing. FAO believe that exposing and sharing terminology resources is a key component in the strategy for achieving these objectives.

The registry provides basic metadata about each KOS, including the name of the resource, the type of KOS (based on a taxonomy developed by the Networked Knowledge Organization Systems Working Group), version and update dates, contact information and a direct link to the source website for accessing the full KOS. The FAO has been working with other organizations, most notably the International/Interagency Ecoinformatics Collaboration Ecoterm Working Group, to develop web services around the registry, to develop an agreed upon core set of metadata elements and to add other environmental KOS resources to the registry either physically or virtually.

Similarly, Mungal sees an increased demand for service-oriented access to terminologies to support description of and access to resources to which the terminologies have been applied or that naturally contain the terminologies. The metadata needed to describe and register terminology resources were discussed in several papers. Mungal discussed the need for metadata in the context of services on the grid, and in particular the needs of caBIG, the Cancer Bioinformatics Grid. As the number of resources increase, the need for consistent metadata that can be queried at the service level has become more apparent. Use cases were collected, including browsing for an appropriate ontology, determining the most authoritative

resource and comparing resources. Mungal identified the types of metadata needed to satisfy each use case.

Mungal then reviewed and attempted to align several standards related to terminologies and terminology resources, such as SKOS (Simple Knowledge Organization Systems from the W3C); the Dublin Core; ISO 11179 Parts 2, 3 and 6; the Common Terminology Service (including Edition 2); and implementations such as the BioPortal's ontology registry. The project found that there is no single "silver bullet." Many of the necessary metadata elements were found in all the standards with some differences in terminology and slight misalignments, but there was no single standard that supported all the use cases. A combination of approaches and additional attributes would be needed, particularly in the areas of quality, version control and "fit for purpose."

Shared meaning. The Semantic Web relies on the explicit sharing of meaning. Keck observed that the disciplines of data management, vocabulary management and ontology management have historically been distinct, and so their standards have tended to be stove-piped as well. At the same time that large data repositories and grids, such as the DOE Science Grid and the Cancer Bioinformatics Grid (caBIG), are being developed, there is no link between these repositories and the explicit expression of meaning that is needed to make the data useful in the Semantic Web environment. Edition 3 of 11179 enhances the ability to relate data elements to concepts, thereby giving data more explicit meaning. Now in committee draft, 11179 Edition 3 elaborates on the concept model, which appears in Edition 2, adds ontology registration and supports the ability to reason over the data using the concept system. ISO 11179 Edition 3 will take advantage of the rich semantics that thesauri and ontology structures can bring to a registry.

The ability to share meaning by more closely tying semantics to data elements in the registry is being piloted through the eXtended Metadata Registries Project [Bargmeyer]. The pilot reference implementation is using sample metadata registry content and complementary concept systems including thesauri, taxonomies and ontologies to test the enhanced model for Edition 3. Use cases include reasoning across the metadata and comparing and harmonizing metadata schemes based on the shared meaning provided by the concept system.

Harmonization and crosswalks. In the process of developing 11179 Edition 3, the ISO/IEC SC32 WG2 is working closely with representatives of TC37 and 46 to align the data models and to crosswalk or harmonize the terminology used among the standards. This effort will be aided by the development of data categories, or metadata for describing each term, under ISO 12620. [Pozzi Pardo]. The harmonization of TC37 standards and the alignment of key concepts will support interfaces between metadata registries and the terminology or ontology systems that provide concepts for classifying the elements. It will also allow metadata registries to more reliably import concept system structures.

Similarly, Reece identified similarities and differences between ISO/IEC 11179 and the ISO 191XX family of geographic information standards. ISO 19115 – Geographic Information and ISO 19135 – Geographic Information-Procedures for Item Registration both drew early on from the 11179 family of standards. However, in the meantime, 11179 continued to evolve to meet new needs and reflect changing practices. The result is that the 191XX family is very specific to the geographic subject matter and 11179 is a more general standard. Both deal with registration, and they share some attributes. However, the relationship between the two is not clear. It is important to bring the current practices among these standards together so that the specifications can be harmonized, not by subject matter, but by management practice. At the very least, a crosswalk between the standards is needed that addresses attributes and registration.

Metamodels. A major method for achieving interoperability is represented by the work on the Meta-Object Facility (MoF) [Horiuchi]. This is a standard that describes how models, primarily those expressed in UML (Unified Modeling Language), can be interfaced. Examples were presented that use MoF and extensions to interface models within an enterprise. It is also considered as a viable means of interfacing the various standards models, such as those for metadata registries, controlled terminologies and specific applications areas.

Horiuchi reported on the trial use of ISO 19763, the Metamodel Framework for Interoperability (MFI), which is based on MoF, to cross industry business practices. The goal is the interoperability between

HODGE, continued

heterogeneous domain registries, such as manufacturing, logistics and retail, through a Registry of Registries. Today, many registries are actively being used in a variety of business domains. However, they were developed to reflect the needs of the specific domain, with its own structure and procedures and with little concern for other registries. As the need for information sharing increases, it now becomes more important to bridge these registries.

The Registry of Registries uses the MFI metamodel for registering process models, ontologies and model mappings. Current discussion topics include registration procedures, how to map models, how to specify data quality and how to provide universal identifiers where needed. Another key issue is the connection between MFI and ISO 11179 Metadata Registries. Horiuchi proposes registration of administration information between the two standards and a common registration and maintenance procedure.

The MFI is also being used to promote access to and re-use of domain models represented as ontologies. ISO 19763-3, Edition 1 provided a standard for a metamodel for ontology registration. However, ontologies evolve over time as the domain being represented changes. Therefore, 19763-3, Edition 2 supports the mapping and comparison of ontologies [He]. The current working draft focuses on the basic model, the types of changes that can occur as ontologies evolve and what information should be kept about the evolution process. Future work will focus on other challenges such as aligning ontology versions. The researchers at Wuhan University will continue to apply and ground-proof the evaluation information model using specific projects in China.

A common framework and transforms. Piprani and Chapin [Piprani] propose the use of a common framework to connect business semantics and IT systems. Business people speak of concepts and representations which are in people's minds. Information systems speak of semantic metadata (usually expressed as data elements and values) and metadata. Data models are concerned with how the data is recorded and stored as part of the information system design. Piprani recommends the use of Structured

Business Vocabulary and Rules (SBVR), ISO 19763, the ISO 704 and 1084 terminology standards and 11179 to connect the language of an enterprise's business people to its IT staff and systems.

A series of transforms are specified. The process begins with the identification and definition of key business concepts using approaches contained in the TC37 terminology standards. These concepts may be further organized into taxonomies or thesauri using appropriate TC46 standards. The concepts are then used in the SBVR business rules that are stated in a business ontology using a structured language. The concepts also become administered items in a 11179 Metadata Registry. The business rules are transformed into constraints in the data model.

The key to successfully bridging the terminology and information science cultures is a clear definition of the key business concepts and keeping the business rules independent of the specific process or workflow. The processes should be independent of the IT and software platform and how they are converted to the models needed by the information systems.

Continuing the Discussion

While the Open Forum provided an opportunity for standards developers to provide updates on their activities and for practitioners to describe their approaches to and ongoing need for the integration of the various standards, this is not the end of the conversation. Discussions continued immediately following the formal conference in a small workshop to address gaps in standards development as posed by the National Cancer Institute's practical use of the various standards in their caBIG development. The discussions may continue through study groups working within the formal ISO/IEC SC32 standards structure. The study groups will include representatives from other standards groups including TC37 and W3C. Discussions will continue within and across these information management standards groups as they seek to improve the interoperability of standards-based systems and the interoperability among standards themselves. ■