

**SOURCE:** TSACC, Telecommunications Standards Advisory Council of Canada

**TITLE:** The Harmonized Use of Metadata in Standards Documents

**AGENDA ITEM:** EWG, 5.5

**DOCUMENT FOR:**

Decision	X
Discussion	X
Information	

## **1 DECISION/ACTION REQUESTED**

*In a box, this should give a very clear statement of what is wanted.  
Proposed formal Resolutions should be drafted in full.*

## **2 REFERENCES**

*(Reference - in list form - should be made to previous GSC/RAST documents or minute references, or to other readily-available sources).*

## **3 RATIONALE**

*(The documents should then set out, as a summary with bullet points, the reasons for the proposed action. The objectives of the proposal should be clearly stated. Other critical considerations should be identified (e.g., strategic, technical, quality, financial, legal, public policy). Rejected alternative solutions should be mentioned if this aids understanding).*

## **4 CONSEQUENCES AND IMPLICATIONS**

*(The implications for (including human resources) should be set out in this section).*

## **5 ISSUES FOR DISCUSSION**

*(This section should, as necessary, contain additional argumentation).*

## Introduction

A commonly used definition of metadata is that it is data about data (1). However, this leaves it as deceptively simple. More precisely, ISO 11179 defines metadata as “the information and documentation which makes data sets understandable and shareable for users” (2). The World Wide Web Consortium (W3C) describes metadata as “machine understandable information about web resources or other things and that it is structured descriptive information about other data; that is, used to aid the identification, description, location, and management of web resources” (5).

Metadata can function in two distinct ways. “One, to provide a means to discover that the data set exists and how it might be obtained or accessed. Two, to indicate how a resource may be used, by documenting the content, quality, and features of a data set “(5). Metadata is comparable to natural language, in that its function is to communicate information about information.

Metadata helps to improve the probability of a client finding a document on-line. Moreover, it has the function of supporting control and management of collections of data (7). It not only aids in finding information it also assists researchers by identifying important components of the information bundle on an object.

## Metadata Registries and Tools

A metadata registry is intended to provide interoperability and extensibility at the global, regional, or domain level in an organization or institution (5). Similar to standards, the registries are organized according to the type of metadata being used. They contain data sets that are commonly agreed within a specific community of interest, maintained by a mutually-recognized authority, and updated under the framework of a common set of requirements by the registration authority. Some example registries are the Australian National Health Information Knowledgebase and the Environmental Data Registry set up by the US Environmental Protection Agency (5).

There are two categories of tools for metadata: editors and generators (15). Editors assist in the creation of metadata by providing a template for entering new metadata content. These editors provide simple means for populating the associated data values. They take the inputs using a user-friendly interface and then translate the content into appropriate formats, such as HTML tags. These translations can be into the appropriate structural location of a document. In the case of HTML, for example, the tags are placed in the HEADER part of a page. Some examples of these are

- Nordic Web Index Dublin Core metadata template
- GEM
- MetaStar Data Entry, and
- Reggie.

Generators, on the other hand, build metadata tags by reverse engineering documents. For example, generators create tags by examining existing HTML-encoded documents and creating the HTML metadata tags. Some examples of these are

- Dublin Core DC-dot
- Medical Metadata project and
- TagGen (15).

## **Metadata Structure**

As a result of extensive work in developing international standards for data sets, the structure of metadata is similar across a wide range of applications and vertical groups. The set of metadata is comprised of data elements that are linked logically by a mutually agreed structure or architecture. The simplest method of connecting sets of metadata is by creating crosswalks or horizontal sectoral mappings. The best example in the context of standards development is the SGML concept of a Document Type Definition (DTD) (5). To achieve more widespread use and heterogeneity, a more complex type of architecture has been developed known as Resource Description Framework (RDF) (5).

In summary, the DTD's are created to specify the meaning of the element within the context of the metadata type. A crosswalk is the conversion from one syntax of metadata into another. Searching across syntaxes and databases is made possible because crosswalks establish a commonality in the definition and the use of the elements (5).

## **Applications**

Currently, Internet merchants who want to have clients find them more effectively are using metadata in a commercial context. Many of the prominent web crawling engines and web search services use metadata editors and generators to create their search data stores. As more use is made of metadata in web documents, then significant improvement in Internet retrieval is possible (5).

## **Standards**

Each special interest sector has developed metadata standards to fit its own needs. For the retrieval of bibliographic or library information on the Internet, the Dublin Core is the most widely used standard. The US government has adopted its own bibliographic metadata standards called GILS. Library cards and libraries most often use MARC, a very early form of defining data about data. Museums have accepted CIMI as their metadata foundation and the geospatial committee relies on FGDC. (3) Internationally, ISO 11179 is the most widely established standard with regards to data.

### **International Standards**

#### **ISO 11179**

Internationally, ISO 11179 is the most recognized standard for the development of commonly shared data sets. The standard consists of six parts: a framework for the specification and standardization of data elements, classification for data elements, basic attributes of data elements, rules and guidelines for the formation of data definitions, and naming and identification procedures for data elements (9).

The first part of the ISO11179 standard contains a glossary of terms, the fundamental model of data elements, a description of the other parts, and an informative index describing the relationship between modeling principles and the standard. The fundamental model divides data elements into three categories: object class, property, and representation.

The second part of the standard provides procedures and techniques for associating data element concepts and data elements with classification schemes for object classes, properties, and representations. In other words, it develops a set of principles, methods, and procedures for specifying what is needed in a

taxonomy/ontology for description of object class, property, representation, and data element concepts and data elements.

The third part is applicable to all of the following activities: definition and specification of the concepts of data elements dictionaries, design and specification of application-oriented data models, databases and messages for data interchange; actual use of data in communications and information processing systems; and interchanging or referencing among various collections of data elements.

The fourth part provides guidance on how to develop unambiguous data element definitions. The fifth looks at identification, which includes the assignment of numerical identifiers that have no inherent meaning to humans and also names, which are semantic, neutral language labels given to data elements. Finally, the last section provides information on how a registration applicant may register a data element with a central registration authority and the allocation of unique identifiers for each data element (9).

## **MARC**

Machine Readable Cards or MARC is the premier cataloguing system used by libraries. It is the standard used to transmit data in the professional library world (12). MARC is the standard for the representation and communication of bibliographic and related information in machine-readable form. In 1973, CAN/MARC was implemented as a distinctive MARC format that would address particular Canadian needs (i.e., bilingualism) (18). A CAN/MARC involves three elements: the record structure, the content designation, and the data content of the record. More specifically, a CAN/MARC format is a set of codes and content designators defined for encoding machine-readable records. These formats are designed for five types of data: bibliographic, holdings, authorities, classifications, and community information. Recently, the CAN/MARC and the USMARC have merged to form the MARC 21. Both countries are making all of their metadata using this standard to ensure greater compatibility (18).

## **CIMI**

The Consortium for the Computer Interchange of Museum Information or CIMI was designed through consensus of participating museums and specifies the following: "a conforming subset of ISO 23950 that must be implemented to ensure interoperability between museum software vendors, the searchable fields in museum databases, and the elements of museum records" (12). It includes a group of institutions and organizations that support an open standards-based approach to the management and delivery of digital museum information (19). They have made a great deal of progress in researching for the museum community standards for structuring its data and for enabling widespread search and retrieval capabilities.

CIMI has a mission statement that clarifies what the true purpose of the organization and the standards are. They encourage open standards-based approaches to creating and sharing digital information. They apply standards to museum information in demonstration projects that invite members to participate and help them to further develop their information systems. They disseminate the results of these collaborative efforts throughout the consortium. And lastly, they share individual work on electronic information issues with each other (19).

## **Dublin Core**

Dublin Core is a 15-element metadata element set intended to facilitate the discovery of electronic resource. These 15 metadata elements are for content, intellectual property, and instantiation (12). At the beginning the Dublin Core initiative was to attempt to define a core set of metadata elements for source discovery. Some of the principles of metadata defined by the Dublin Core are as follows (5):

- metadata takes a variety of forms that corresponds to unique areas of expertise
- new metadata sets will develop as the network information infrastructure matures
- different communities will propose, design, and be responsible for different types of metadata
- there are many users of metadata
- metadata and data have similar behaviours and characteristics
- the metadata sets associated with an object may be physically allocated or reference indirectly.

There are five distinct characteristics of the Dublin Core's project. They are simplicity, semantic interoperability, international consensus, extensibility, and metadata modularity on the web. These are the characteristics that the Dublin Core hoped to accomplish through its project (17). Simplicity means that the metadata is intended to be useable by non-catalogers as well as description specialists. Promoting a common understood set of descriptors that helps us unify other data content standards should increase the possibility of semantic interoperability across disciplines. By having active participation and promotion in over 20 countries, an international scope of resource discovery on the web has been established by the Dublin Core. The extensibility provides an economical alternative to more elaborate description models. It includes sufficient flexibility and extensibility to encode the structure and more elaborate semantics inherent in richer description standards. Finally, the diversity of metadata needs on the Web requires infrastructure that supports the coexistence of complementary, independently maintained metadata packages (17).

## **National**

### **NCITS L8**

Aside for the international body, which designed a standard for metadata, there are also national bodies who have done the same. For example, NCITS L8 (National Committee on Information Technology Standards, Technical Committee L8) establishes standards for specifying and standardizing data (10). Some metadata issues covered by the committee include naming, identification, definitions, classification, and registration. These standards are used in areas like EDI, data administration, information management, and data access/interchange via the World Wide Web and the national information infrastructure (10).

### **GILS**

GILS (Global Information Locator Service) is a standard that makes it easy for people to locate information and evaluate its utility (11). The standard itself specifies how to express a search and return results in all languages (20). It provides you with a description of publicly available information on your topic of interest. Each record presents a thorough description of the information resource. It tells you what information is available and why it was created, how the information is made available for you use, who to contact for further information, and a direct electronic link to the information (12).

The Executive Office of the President, Office of Management and Business, under OMB Bulletin 95-01 established GILS (24). In December of 1994 the GILS profile was approved by the Department of Commerce as a Federal Information Processing Standard, which required compliance by government departments (23).

GILS defines four primary users of the standard and how GILS aids in their search (20). The first is a content owner. GILS helps to assure people can find the information the owner is offering. Intermediaries is the second and is assisted by GILS by making it easier for the intermediary to gather information from any source described. GILS allows a direct user to use intermediaries or to search primary sources directly. Finally, software companies can benefit from GILS because they provide a powerful platform

adaptable to a range of architectures if the company wants to provide search access to information on the desktop, Intranet or Internet.

In 1995, the Treasury Board set up a working group to look at the possibility of accepting GILS into the Canadian government (24). The working group became a subgroup of the Electronic Document Standards Working Group. During its life span the group was to establish guidelines for a Canadian version of GILS, devise a pilot project of GILS, and make revisions to the GILS Application Profile. These activities were all done in order to determine if GILS is acceptable to be a Treasury Board Information technology Standard.

## **FGDC**

The first group in the US to start looking at the standardization of metadata and become the most active in the area was the Federal Geographic Data Committee (FGDC). President Clinton issued an executive order on April 8, 1994 titled "Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure" (21). The Content Standards for Digital Geospatial Metadata was developed and passed on June 8, 1994, in response to the request. It became the federal standard for geospatial data. The standard was developed from the perspective of defining the information required by a prospective user to determine that availability of a set of geospatial data. It was also to determine the fitness of the set of geospatial data for an intended use. The standard was to determine the means of accessing the geospatial data and to successfully transfer the set of geospatial data (21).

The standard specifies the information content of metadata for a set of digital geospatial data. The purpose of the standard was to provide a common set of terminology and definition for concepts related to metadata. It includes such areas as identification information, data quality information, spatial data organization information, spatial reference information, entity and attribute information, distribution information, and metadata reference information (14). The objectives of the standard, and coincidentally the major uses of metadata, are to maintain an organizations' internal investment in geospatial data, to provide information about an organization's data holdings to data catalogues, clearinghouses, brokerages, etc, and to provide information needed to process and interpret data to be received through a transfer from an external source (13).

## **Challenges**

One of the major challenges facing standards bodies in developing information about their respective standards is to provide a mechanism for providing accurate and timely information to those interested in obtaining the standards. Over the years, the Electronic Exchange forum has been examining ways of exchanging standards information among the partners in GSC. We have examined many different schemes, including the use of proprietary systems and commercial packages. Implementation of metadata would allow us to freely exchange information about our holdings and open exposure of our standards documents through standard Internet World Wide Web search facilities. It has proven extremely useful and effective in the bibliographic and library, geographic, and health sectors. As developers and publishers of standards, we have many similar characteristics for data exchange as these other sectors.

The challenge will be to develop a harmonized set of metadata structures that would facilitate the exchange of information on standards across our various organizations. The value of metadata has already been established in other sectors. It is proposed that the GSC partners begin a project to develop a set of common metadata elements to describe our information holdings and facilitate the search of those holdings.

## Glossary of Terms

### **RDF- Resource Description Framework**

The Warwick Working Group developed Resource Description Framework or RDF and presented it to W3C for approval in February 1998 (16). RDF is an infrastructure for web metadata. It is a uniform and interoperable means to exchange metadata between programs and across the web. RDF is a domain-neutral knowledge representation mechanism. It can be used for resource discovery, cataloguing resources, intelligent software agents, content rating, creating and describing collections, for "intellectual property" rights, and with digital signatures to create a "web of trust" (16).

### **Crosswalk**

A crosswalk is the accurate conversion of one syntax of metadata to another syntax. Searching across syntax's and databases is encoded because crosswalks establish a commonality in the definition and the use of the element. A crosswalk was created between the Dublin Core, MARC, and GILS that provided a very specific MARC record for formatting information. This was done by the Library of Congress.

### **DTD – Document Type Definition**

It was created to specify the meaning of the elements within the context of the metadata type. Technically, it is the allowed structure and combination of structures within a document. DTD is analogous to schema, machine-readable definitional structures or relational databases.

### **XML – Extended Markup Language**

To work hand-in-hand with RDF is XML. XML is a form of metadata. Its construction allows SGML to be delivered over the web, and as a result, overcomes the limits of HTML. It has the ability to render broadly functional and valuable business applications on the Internet, Intranets, and Extranets. It is driven by applications like RDF (5).

### **SGML – Standard Generalized Markup Language**

SGML has been described as an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form (22). It is a metalanguage or a means of formally describing a language, in this case, a markup language.

SGML is different from other markup languages for three reasons. One, it places a greater emphasis on descriptive rather than procedural markup. It uses a document type concept and it is independent of any one system for representing the script in which a text is written (22).

### **Markup**

A markup describes an annotation or other marks within a text intended to instruct a compositor or typist how a particular passage should be printed or laid out (22).

## **Markup Language**

A set of markup conventions used together for encoding texts. It must specify what markup is allowed, what markup is required, how markup is to be distinguished from texts, and what the markup means (22).

**References**

1. Day, Michael & Powell, Andy (1998). UKOLN Metadata homepage. <http://www.ukoln.ac.uk/metadata> .
2. Newton, Judith (1996). Applications of Metadata Standards. NIST. <http://www.computer.org/conferen/meta96/newton/paper.html>
3. Blue Angel Technologies (1999). FAQ: Metadata Explained. <http://www.blueangeltch.com/NewsAndInfo/FAQs/metadata.htm>
4. <http://www.ukoln.ac.uk/metadata/desire/overview/revti.htm>
5. Hodgson, Katrina (1998). Metadata: Foundations, Potential, and Applications. [http://www.slis.ualberta.ca/538/khodgson/metadata.htm#what\\_is\\_metadata](http://www.slis.ualberta.ca/538/khodgson/metadata.htm#what_is_metadata)
6. Interoperability Committee (1999). What Every CIO Needs to Know about Metadata. <http://www.itpolicy.gsa.gov/mke/archplus/first.htm>
7. Hakala, Juha; Husby, Ole; & Koch, Traugott (1996). Report from Metadata Workshop II on Warwick Framework. <http://www.ub2.lu.se/tk/warwick.html#2>
8. Rzepa, H.S. (1996). Chemical Metadata Standards for the World Wide Web. <http://www.ch.ic.ac.uk/chemime/chemeta.html>
9. Skall, M.W. (1999). Overview of ISO/IEC 11179, Parts 1-6. NIST. <http://sdct-sdct-sunsv1.ncsl.nist.gov/~ftp/18/other/coalition/ovr11179.html>
10. McCarthy, J.L.; Olkin, F. & Perelman, N. (1998). NCITS L8- National Committee on Information Technology Standards, Technical Committee L8. <http://www.lbl.gov/~olken/X3L8/L8intro.html>
11. ??? <http://iep.fedworld.gov/library/elapbiggs/appendixa.html>
12. Blue Angel Technologies (1999). FAQ: Metadata Explained. <http://www.blueangeltch.com/NewsAndInfo/FAQs/glossary.htm#GILS>
13. Schweitzer, P. (1998). Content Standards for Digital Geospatial Metadata. <http://geology.usgs.gov/tools/metadata/standard/overview.htm>
14. FGDC (1998). Geospatial Metadata. NSDI – National Spatial Data Infrastructure. Metafact. <http://www.fgdc.gov/publications/documents/metadata.metafact.pdf>
15. National Library of Australia (1999). Meta Matters – Tools. <http://www.nla.gov.au/meta/tools/html>
16. Rew, R. (1998). Metadata and RDF Survey. <http://www.unidata.ucar.edu/staff/russ/bs/metadata/index.html>
17. Dublin Core (1999). Dublin Core Metadata Initiative. <http://www.purl.org/DC/>
18. National Library of Canada. (1998). Introduction to CAN/MARC. <http://www.nlc-bnc.ca/marc/eintro.htm>
19. CIMI Homepage (1999). Introduction. <http://www.cimi.org/about/introduction.html>
20. Christian, E (1998). Global Information Locator Service. <http://www.gils.net/locator.html#find>.

21. FGDC (1999). FGDC Metadata. <http://www.fgdc.gov/metadata.html>
22. TEI (1998). A Gentle Introduction to SGML. <http://www.oasis-open.org/cover/gentle.html>.
23. Government of Canada (1997). GILS Overview. <http://gils.gc.ca/overview/overview.html>
24. Government of Canada (1998). Canadian GILS Guidelines.  
[http://gils.gc.ca/english/gils\\_info/guide\\_e/htm](http://gils.gc.ca/english/gils_info/guide_e/htm)

Web-sites of Interest

<http://www.its.nbs.gov/nbs/meta/meta.htm>

<http://linnea.helsinki.fi/metadata.nmfinal.htm>