Assuring Semantic Consistency for Data Interchange

Posted on January 1, 2004 by Judith Newton

Published in TDAN.com January 2004
The adoption of XML as the data interchange format for the Web presents a set of challenges and opportunities for data managers. While XML makes it easy to describe the format of information objects and the relationships among them, it does nothing to assure their semantic consistency. Supplementing XML schema descriptions with some mechanism to document the metadata helps determine the meaning of each object in relation to similar objects. Past problems, and their solutions, in data management have shown that known semantic consistency is the key to successful information interchange.

XML has had a tremendously positive impact on the connectivity of systems, but also has more clearly exposed what problems remain. XML is a markup language that can be used to tag data elements and collections of data with labels. As part of a standardization activity, communities can agree on the names for these labels. Problems arise, though, if different users have differing understandings of the meaning of an XML tag. In other words, XML standardizes the syntax of data exchange, but was never designed to capture the meaning of the data. This is not necessarily an obstacle for a community that operates in a common context; there, the mental associations with a tag are shared and well understood by all. Where this limitation becomes a problem is in moving data from one context to another, for example, in sending data from a manufacturing context to a financial context. Without explicit, rigorous definitions of terms, misunderstanding is sure to arise.

One way to associate meaning with XML elements and attributes is through linkage to a metadata registry (MDR). A metadata registry can be used to store names and descriptions of data units and data values, and information about their organization and representation in many applications. As such, it can assist with issues of naming and identification, metadata description and organization for XML artifacts. It can be used to store XML names for entities, and to associate them with explicit definitions and other descriptive information. In particular, in developing MDR's, principles have been developed for the establishment of standardized names through documented naming conventions, and these conventions can also be applied to XML names, and to sets of XML names ("XML namespaces").

Where Do Namespaces Come From?

Many efforts have arisen to address development of standardized XML schemas peculiar to specific subject areas. Within these areas one or more **namespaces** define sets of metadata elements (MDE's) within schemas. These efforts open opportunities for data stewardship and for managing data semantics, including naming issues, over a broad range independent of the physical systems which process the data.

Traditionally, data managers have used namespaces to divide their enterprise or area of interest into manageable chunks; these may be thought of as "functional" namespaces. Business rules have been applied within namespaces. One of these rules has been a determination of uniqueness, not just of names, but also of metadata entity content within namespaces. Rules governing relationships among namespaces have also been applied.

The history of data management has been one of "islands" of well-formed data – including metadata – developed without coordination or integration. These islands arose for various reasons:

- A set of well-documented metadata was supplied with an off-the-shelf tool or (more likely) was the deliverable specified as part of a consultant's contract
- An advocate arose with a limited but effective sphere of influence
- A higher authority forced the system developers to comply with organizational standards

The problem of bridging these islands seemed insurmountable in the days of batch updates and file transfers by physical media. In the 1980s, academicians struggled with "federated schemas" for distributed databases, but in the 1990s, many businesses moved to a single huge integrated schema, typically supplied by a single vendor, and replaced all their applications to work with it. With new technology, however, come new opportunities.

Making the links between consortium-developed XML name sets by which an application must communicate, and enterprise-developed name sets used in the databases raises a new and interesting set of issues. These can be addressed by the adaptation of the knowledge gained by hard experience in the last thirty years. Let's examine some of those principles, particularly those concerning names, namespaces and metadata.

Names

Understanding the semantics of names is part of a broader issue of getting computers to parse linguistic connections. The proliferation of names has many causes. Just as each application of data has a unique set of requirements and restrictions for the names used in that application, each new technology creates requirements for the use of names and constrains the names that can be conveniently used. XML is such a technology, and it places heavy emphasis on names.

A metadata registry allows each of its metadata elements (MDE's) to have any number of names. One or more of these names should have meaning to the user, by describing the content of the element in a structured way, derived by a rigorous process, using a formalized naming convention.

MDE names collected in the MDR may occur in any context. For example, these may be:

- Software system names
- Programming language names
- Report header names
- Data interchange names

They may have varying levels of rigor applied to their formation and usage. The capture and display of all names used by any one MDE is a major strength of the MDR model. Adding XML names to the set of names documented for MDE's minimizes the confusion. It captures the relationship between the XML name and other names for the same MDE used in programs and databases, and it documents the source of an XML element when the XML schema is generated from a set of existing MDE's (such as SQL schemas or databases).

A structured way to form names, which derives from the meaning of the MDE itself, utilizes semantic components from the MDE definition or structure sets (data models, taxonomies, etc.), of which it is a member. Names from other contexts may then be documented in the same entry.

Naming Metadata

The units used to form names are described by ISO 704, *Terminology work – Principles and methods* [ISO 704] and ISO 1087-1, *Terminology work – Vocabulary – Part 1: Theory and application* [ISO 108] as follows:

- A **term** is a verbal designation of a general concept in a specific subject field.
- An **appellation** is a verbal designation of an individual concept.
- A **symbol** is a visual representation of a concept.

A structured name is derived from a naming convention. Naming conventions consist of sets of rules. These may include semantic, syntactic, lexical, and other types of rules according to the degree of rigor required. Terms, appellations and symbols serve as the semantic components of names, and semantic rules describing the derivation of these components form the semantic rule set of a naming convention. The order of components within a name is determined by the syntactic rule set. Lexical rules are applied to determine the appearance of components.

Here is an example of a semantic rule set using the concept system for metadata element components described by the MDR metamodel. The components of this system are:

- Object Class – a set of ideas, abstractions or things in the real world
- Property – a characteristic common to all members of an object class
- Representation – describes the depiction of the values in the value domain of the MDE

These components can be transformed into terms forming the semantic component of an MDE name. They become object class terms, property terms, and representation terms. They are all general concepts; as such, they all may be named by terms, but to allow for greater flexibility among words in a name, appellations are also allowed. As 'representation' is the name component that

documents the logical form in which the data element is presented, representation is the only name component in which symbols are allowed.

An example of a rule set for the semantic part of a naming convention:

- Object class terms shall be derived from the names of object classes.
- They shall consist of: at least one term, and zero or one appellation.
- One and only one object class term shall be present.

- Property terms shall be derived from the names of the property terms associated with the object class term in the model.
- They shall consist of: at least one term, and zero or one appellation.
- One and only one property term shall be present.

- Representation terms shall be derived from the names of representation classes.
- They shall consist of: one term, or one symbol, or a combination of one term and one symbol.
- One and only one representation term shall be present.

In addition to these rules, rules for modifier (qualifier) terms and significance of separators may be present.

Metadata Registry

ISO/IEC 11179, *Information technology – Metadata Registries*, [ISO 111] is a six-part standard describing a conceptual model for collecting and organizing metadata. The semantic
information contained may be collected from anywhere in an enterprise's area of interest. The standard does not specify any particular implementation; the registry may be an independent product,
incorporated into an existing product such as a data repository, or other system architectures as desired.

Using a metadata registry based on ISO/IEC 11179, users can store metadata about the classification, naming, identification, definition, and organization of information in order to make it
understandable and shareable. Data about sources, usages, and derivation of information are made available in a readily accessible form. Also, the rules for registering and defining information
units, along with other conventions, are documented.

Using a conceptual metamodel allows relationships among differing representations and value sets of the same information to be mapped together in one place. This is useful for tracking the source
of the XML objects generated for interchange back to the original usage, and documentation of other usages of that information within an organization.

Some other documentation capabilities that are available in the MDR include:

- Documentation of data structure through classification. The MDR provides a Classification region, in which the structure of data can be described. Namespaces are one example of a structure that
  can be documented in the classification region and linked to tags recorded for each MDE.
- Data stewardship information. The MDR has extensive provisions for documentation of the stewardship contact information for each MDE entry.
- Versioning capability. Every MDE has a built-in versioning mechanism.
- Visibility and Understandability. Linking an XML structure to an MDR-based registry makes additional benefits available to XML tools.
- Promotion of interoperability. Interfaces can be documented in the MDR and made visible to users.
- Trustworthiness assessment. The MDR can provide documentation for sourcing, timeliness, collection methods, and other means of confidence assurance.

One part of an MDR of particular interest to XML users is the **value domain**. A value domain is the set of potentially valid values for one or more MDE's. It is used for validation
of data in information systems and in data exchange. It is also an integral part of the metadata needed to describe a data element. In particular, a value domain is a guide to the content, form,
and structure of the data represented by a data element. A non-enumerated value domain may be described by definition, reference, or rule. An enumerated value domain is defined by a list.

The equivalent concept to an enumerated value domain of an MDR in an XML schema is an enumerated list (properly, a restriction of a simple type to a set of 'value' facets), used to document the
possible valid values in a domain. It is the mechanism used for listing code values. However, domains with more than just a few valid values are difficult to describe within the schema, and many
code lists have hundreds of valid values. A link from an XML schema to an MDR means that the schema no longer needs to carry the code values.

Namespaces in the MDR

XML developers group names (and declarations) into schemas or DTD's for particular business requirements. Each schema or DTD usually defines all the names in one XML namespace (and may include names from other namespaces). These names may have been developed using any number of naming conventions, including none. There may also be a need to stipulate that only specified naming conventions are utilized for a particular namespace.

The relationships of namespaces, naming conventions and names must be documented within a registry if names and the use of names produced by naming conventions are to be detailed for business requirements. The current version of the MDR can support namespaces in several configurations. The decision of how to document namespaces in the MDR should be based on how much metadata must be recorded on the namespaces themselves, in addition to their content.

An MDR can assist XML users in maintaining the link between XML components and their sources, and in retaining, and providing access to, extensive knowledge about the data, by storing metadata that would make XML structures unwieldy. Meaning is maintained by using semantic components to form names; by using conventions within namespaces; and by using an MDR as a rich metadata resource to augment the sparse metadata descriptive mechanisms XML provides.

References

[ISO 704] ISO 704:2000, *Terminology work – Principles and methods*, International Organization for Standardization, Geneva.

[ISO 108] ISO 1087-1:2000, *Terminology work – Vocabulary – Part 1: Theory and application*, International Organization for Standardization, Geneva.

[ISO 111] ISO/IEC 11179, *Information technology – Specification and standardization of data elements*, Parts 1, 2, 4, 5, 6, International Organization for Standardization, Geneva.

[ISO 111] ISO/IEC 11179:2003, *Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes*, International Organization for Standardization, Geneva.

Note: ISO/IEC 11179 is a six-part standard currently undergoing revision and a change of title. Part 3 is the first Part to be published in the second edition.

Share

Share reddit this!

About Judith Newton

Judith is Principal of Ashton Computing and Management Services, LLC, a consulting firm specializing in web design and metadata development. She is currently the Senior Analyst for two metadata registry development projects.

She is a U.S. delegate to the International Standards Organization Subcommittee for Data Management and Interchange (ISO/IEC JTC 1/SC 32), Working Group 2, Metadata, and author and editor of the ISO Standard on *Metadata Registries: Naming and Identification Principles* (ISO/IEC 11179-5) and the technical report *Specification of Data Value Domains* (ISO/IEC TR 15452). She is editor of the technical report on *Procedures for Achieving Metadata Registry Content Consistency: Data Elements* (ISO/IEC PDTR 20943-1).

She is a member of ANSI INCITS L8, Metadata, which is U.S. TAG to SC 32/WG 2. As Chair of the L8 Task Group for Technical Development, she led the technical development and consensus process to achieve completion of products at the national and International level.

Ms. Newton is a past member of the American National Standards Accredited Committee for Information Resource Dictionary System (X3H4). In 1992 she chaired the Task Group that Produced the Technical Report *IRDS Support for Naming Convention Verification* (ANSI X3/TR-11-92), addressing the feasibility of an automated naming tool for the IRDS.

Judith was employed by the National Institute of Standards and Technology 1979 to 2004. At NIST, her most recent project involved study of the synergy between XML registries and 11179-based metadata registries. Other projects have addressed enterprise data modeling, data repositories, and semantic interoperability. In a consultant capacity, she has advised several agencies and Federal committees on metadata usage, among them EPA, DoD (DISA), and Navy.

She has also served as president of the Data Administration Management Association (DAMA) National Capital Region Chapter (DAMA-NCR), from its founding in 1987 to 1990; and chaired the highly successful DAMA Symposia in 1988, 1989, 1990 and 2001. She continues to serve on the Executive Board of DAMA-NCR. She served on the Program Committee for the DAMA-International/Metadata Symposium 2000, and the DAMA-NCR Symposium 2003.

From 1973 to 1979, she was employed by Navy Regional Data Automation Command (NARDAC), Washington, D.C. to develop and maintain the RAS STADES system, an early effort to manage standard data elements using a data element dictionary system.

She was the recipient of the 2001 InterNational Committee for Information Technology Standards (INCITS) Merit Award, and the 2005 DAMA-International Government Award.

She is a graduate of Temple University.