

VIUS Reports 5.1

Metadata Work

While we examined most available metadata standards designed for visual resources, we quickly focused on three that had particular relevance for the Penn State environment. These were the Visual Resources Association Core Categories (VRA Core), Dublin Core, and an emerging standard, Instructional Management Systems (IMS) Learning Resource Metadata Specification. Because of the broad disciplinary spectrum, if for no other reason, choosing a single standard seemed less important than developing an approach to standards that ensures coordinated searching and portability.

The VRA Core is a significant standard for describing pictures of art and architecture and is rich in data elements relevant to these disciplines. While not universally applied, VRA Core has found many applications. This scheme is significant to the Penn State environment because the largest picture collection, the Art History Slide Library has adopted a modified form of VRA for cataloging their collection.

Because of recent increases in the use of learning management software, the IMS specification was an important new standard to consider. It is designed for the description of learning objects from curricular materials to course-related multimedia objects. Most learning management systems are accommodating the specification and awareness of IMS in the digital library community is growing. The development of projects such as MERLOT (The Multimedia Educational Resource for Learning and Online Teaching) and CICERO (CIC Educational Resources Online) herald an increasingly important role for IMS in academic information systems. Several Penn State projects are employing this standard.

Dublin Core supports resource discovery across disciplines through a simple list of fifteen common data elements. It has proven versatile in promoting interoperability across interdisciplinary collections employing other metadata standards, because many of these standards have been mapped to Dublin Core, if not to each other. Dublin Core thus forms a useful least-common-denominator language. However, such utility is not without cost, as such data mapping involves significant loss of specificity when the richer data captured in the more complex standards are reduced to the more rudimentary Dublin Core elements. This standard is relevant to the Penn State environment because both of the above standards have been mapped to it and because it has been used for previous projects in which the Penn State Libraries have participated.

Initial efforts to map the three standards led us to the realization that element definitions between the standards are not equivalent. (See VIUS Reports 5.2 VIUS Crosswalk) The manipulation of data between elements in a mapping would not be sufficient to make the data truly interoperable. Additionally, the structural differences between IMS, an extremely hierarchical standard, and the relatively flat structure of the others were

glaring. An element's place in a hierarchical structure serves to express some of its character. We simply had no way to avoid losing that element's context when we tried to extricate it in order to map it to another standard. Beyond the structural issues involving data elements, there remain difficult issues involving the data content, terminology, and semantics.

In an effort to support the three metadata standards with no loss of context or hierarchical relationships, members of the project team developed a merged superset of the three. The intent was that the resulting schema, expressed as an XML DTD, could support data transformations for importing and exporting data and for creating a common query language. An ideal VIUS schema would offer a set of elements that corresponds to each element in the three standards, thus allowing us to preserve the specificity of any imported or exported metadata. (See VIUS Reports 5.3 Initial Plan for VIUS Data Element Definitions)

Ultimately this work was abandoned for three reasons: 1) some portions of the problem were unresolvable, 2) the choice of software for our prototyping work prevented deployment of this type of schema, and 3) the database prototyping experiment was focused upon issues of identifying and obtaining content. The choice of software for the prototype database (CONTENTdm™, discussed below) prevented testing our schema in practice. The CONTENTdm™ software supports customized schema for each collection, but requires that each element be mapped to one of the Dublin Core elements for cross-collection searching. Our prototype was bound to the simplicity of Dublin Core. At the end of the project, the thorny issues of equivalent definitions, semantics, discipline specific thesauri, and interoperability with IMS are left unresolved.

Since speed in identifying and obtaining database content was and will always be a priority, existing metadata from several sources (vendors, faculty, licensed images, free digital archives) was obtained and mapped into a simplified Dublin Core. This, then, became the essential task of metadata design. The problem of inheriting complex metadata with an object, pushing it into a DC filter, and losing that information forever, is something we learned to live with. At the opposite end of the spectrum, the problem of having little to no metadata accompanying digital images raised many questions. Given the great variety of data received, we wondered if the exigencies of digital library endeavors mitigate against quality "cataloging," specifically adding authority control and adequate search terms to aid in discovery? Metadata for images is crucial because images can not be mined in the same way as text. The mediation involved in such a large scale application is challenging.

We learned from the user study that faculty do not apply descriptive metadata to their personal collections. Nor do they indicate that they are looking for extensive search terms when it comes to digital images. Particularly in focus groups the question arose as to how little metadata is required to still facilitate discovery. This is an avenue that will be pursued in the development of the peer-to-peer system, LionShare, as we study more about user contributed metadata. Does dynamically added content mean unmediated, uncontrolled metadata? Will users recognize the importance of metadata for discovery?

We allowed some digital collections to preserve their native metadata by mapping to Dublin Core, such as the Campus Buildings collections and the Prints of Pennsylvania collection. For other collections, we scripted incoming data to merge into predetermined field names that would accommodate various subjects and genres. These were then mapped behind the scenes to Dublin Core to allow cross collection searching. A team worked on a clean up effort when we split the collections into two parts, based on permissions: PSU only (private collections) and Public Collections. As part of the clean up effort, statements regarding rights and permissions were added. The field "Category" was created and populated with terms from the Getty Art and Architecture Thesaurus (AAT) to describe the functionality of the buildings, and to accomplish federated searches.

Ann Copeland, John Attig, Michael Pelican, Henry Pisciotta
10/13/03