

# Enhancing Interoperability between Digital Libraries and Educational Technology via XML Crosswalks

A Proposal to the Andrew W. Mellon Foundation from the Interactive University Project,  
University of California, Berkeley

Prepared by Raymond Yee, Ph.D. ([yee@uclink.berkeley.edu](mailto:yee@uclink.berkeley.edu)), Technology Architect and  
David A. Greenbaum ([dag@uclink.berkeley.edu](mailto:dag@uclink.berkeley.edu)), Director, Interactive University  
Project, Information Systems and Technology

## Abstract

Lack of interoperability among software systems and repositories from different domains is a major barrier to the exchange of digital content between communities. This project will explore how semantic interoperability (the accurate translation of meaning) in the following four domains can be enhanced through the use of XSLT-based crosswalks between key XML specifications: 1) digital libraries and repositories (METS); 2) educational technologies and learning management systems (SCORM, IMS-Content Packaging (IMS-CP), and IMS-Metadata (IMS-MD)); 3) web syndication and portal technologies (RSS); and 4) desktop applications and structured content authoring tools. (e.g., Microsoft Office 11).

The deliverables for this project are:

1. Documented and refined crosswalks among METS, SCORM/IMS-CP, and RSS that will serve some immediate practical needs of interoperability among digital libraries, educational technologies, and web syndication systems
2. Exploratory architecture and software prototypes for the deployment of crosswalks
3. White papers summarizing our crosswalk investigations and relating the crosswalks to other approaches to semantic interoperability

We will carry out this work in close collaboration with an Advisory Board of domain experts from the library and educational technology communities.

This project will leverage the work of other collaborations in which the Interactive University Project (IU) is involved. For example, we will soon commence a project with the California Digital Library (CDL) to explore how the resources of the CDL can be more effectively exposed via second generation web techniques. Moreover, we expect to find substantial synergy between this project and our existing collaborations with faculty, research units, libraries, and museums on the Berkeley campus.

## Background

Content that resides in libraries and digital repositories can be of great use in educational contexts. XML interoperability specifications proposed within the library/digital repository community (e.g., METS) and within the educational technology domain (IMS, SCORM) are geared to the exchange of content within that specific community. By

themselves, these standards do not address the need for cross-community transmission of data.

The UC Berkeley Interactive University Project has focused on how content encoded in a particular XML specification can be translated to another XML specification through crosswalks. We have done preliminary work with XML specifications from the digital library and educational technology communities -- as well as those used in the syndication of web content and mainstream desktop productivity tools.

The mission of the Interactive University Project is to use the Internet to open UC Berkeley's unique resources and people to the public, especially California's K-12 schools and citizens. The IU aims to engage the academic core of the campus -- faculty, academic departments, organized research units, libraries, and research museums -- in using technology to structure content so that it can add value to teaching, research, and public service. A key means of structuring content will be to create and use XML-based digital objects. The IU is building the Berkeley Open Learning Environment (B-OLE) for the sharing and creating of such digital objects both on and off the campus.

A key component of the B-OLE is the Scholar's Box that will enable faculty, students, and the public to create, manipulate, annotate, and share personal collections of digital cultural objects gathered from multiple digital repositories -- core activities in both scholarship and teaching. In creating ideas, developing presentations, sifting through evidence, researching papers, or compiling readers, scholars build *de facto* collections from which they create their desired product. Gathering, manipulating, organizing, annotating, and sharing personal collections of cultural objects is also a core activity that can support many teaching and learning practices and styles.

Ideally, the Scholar's Box would enable users to draw upon multiple sources in seamless, integrated ways regardless of underlying protocols and data/metadata encoding schemes. Creating the full spectrum of interoperability required for such functionality remains an extremely challenging and multifaceted research problem. [5] Among the various aspects of interoperability, the problem of semantic interoperability, "integrating resources that were developed using different vocabularies and different perspectives on the data" [3], has been of special interest to the IU.

There have been a variety of attempts to solve the general semantic interoperability problem through the creation of an abstract scheme in which specific vocabularies can be subsumed as particular cases of the scheme. Translation between any two vocabularies is then handled by using the abstract scheme as an intermediary. That is, a translation from a given specification X to specification Y is accomplished by translating specification X to the abstract scheme and then from the abstract scheme to specification Y. The existence of such an abstract scheme renders unnecessary direct translations between specifications, which would grow rapidly in number as the number of specifications increases. [2-4] Of course, constructing an abstract scheme that can accurately subsume all specific vocabularies of interest remains a major unsolved challenge. Nevertheless, projects that aim to solve the semantic interoperability problem in the large are valuable since the desire to work seamlessly with the multiplicity of digital content in varied formats will continue to grow.

Meanwhile, the IU has been pursuing a pragmatic approach to enhancing semantic interoperability among libraries, educational technology, and web syndication. By focusing on a small number of XML-based interoperability specifications that are of importance in the various domains, we have written direct crosswalks between the specifications, thereby avoiding the need for an abstract scheme. In our work, we have produced baseline translations, rather than crosswalks of the highest fidelity. We have transported materials between library repositories and instructional technology applications via these functional converters for three purposes: 1) to demonstrate the interchange of digital content between libraries and instructional technology systems; 2) to learn from practitioners in the educational and library communities where to invest further effort in making the crosswalks useful in production services; and 3) to encourage the developers of interoperability specifications in the library and educational technology communities to harmonize related specifications where possible, thus reducing the need for crosswalks in the first place. This project will enable the IU to make further progress on these three fronts.

The Interactive University Project is deeply committed to this project because enhancing semantic interoperability is crucial to the functioning of the Scholar's Box -- a high priority for the IU. Although the IU is focused on leveraging practical techniques to enhance interoperability, we also integrate insights from long-range research efforts. Working in dialog with experts from the library and educational technology domains helps us develop useful crosswalks. The IU already has partnerships on campus (with the University Library, Berkeley Art Museum, Educational Technology Services, Berkeley Natural History Museums, the Multimedia Authoring Center for Teaching in Anthropology) and off-campus (the California Digital Library) that will help in finding those experts.

The two members of the IU staff leading this project are well qualified for the task. Raymond Yee has the technical knowledge and the educational and scholarly vision to carry out this project. He has been involved in software development for over 15 years. He received a Ph.D. in Biophysics at the University of California, Berkeley in 1997 and a B.A.Sc. in Engineering Science from the University of Toronto. While earning his Ph.D., he also taught K-11 students in the Academic Talent Development Program on the Berkeley campus.

David A. Greenbaum is one of the creators of the Interactive University Project at Berkeley. He has worked for seven years investigating how the campus can best use the Internet to open its research content and community in support of K-12 teachers, learners, and the public at large. Greenbaum's leadership in these efforts has been recognized by the U.S. Department of Commerce Technology Opportunity Program as one of three model projects nationwide for its evaluation report and methods and most recently by Educause, through the Award for Exemplary Practices in Information Technology Solutions. He graduated summa cum laude in Political Economy of Industrial Societies and carried out doctoral studies in Jurisprudence and Social Policy at UC Berkeley.

The IU is a unit in the campus's Information Systems and Technology Division, which is led by Associate Vice Chancellor and campus Chief Information Officer Jack McCredie. The Principal Investigator for the IU (and for this proposal) is the campus's Executive

Vice Chancellor and Provost, Paul Gray. EVC Gray and AVC McCredie will serve as the project's overall leaders, ensuring campus support and guidance for the project.

## Proposed Work

*We propose to 1) refine and publish crosswalks that translate among METS, IMS-CP, and RSS, 2) prototype an architectural generalization of these crosswalks, and 3) connect the crosswalks to other architectural approaches to semantic interoperability.*

### **1. Documentation, Refinement, and Dissemination of Crosswalks**

*We will*

- *publish v. 1 of XSLT crosswalks that translate among METS, IMS-CP, and RSS*
- *document the rationale, logic, limitations, and areas needing refinement*
- *solicit feedback from the Advisory Board on how the crosswalks can be made most useful*
- *iterate to produce v. 2 of the crosswalks.*

We currently have written preliminary crosswalks among METS, IMS-CP, and RSS (Version 1). The crosswalks enable materials to be moved from one environment to another. Structural elements of the materials are preserved whenever possible. A start has been made at metadata translation. In some cases, the crosswalks capture the one and only appropriate translation. In other cases, a choice was made from among a number of possible reasonable choices.

Version 2 of the crosswalks will be refinements based on the input of the Advisory Board (see below for a description of the Advisory Board). The Advisory Board will help us to understand properly the semantics of specifications from the various domains and specifically the nuances around translating concepts from one domain to another.

Moreover, we need to understand the specific contexts in which the crosswalks will be deployed. Since there is often more than one viable crosswalk, we will document the reasoning behind the choices we make so that the crosswalks can be intelligently recontextualized as needed. The Advisory Board will help us determine important practical scenarios to address.

The Interactive University has been collaborating with the California Digital Library (CDL) in exploring and prototyping how the content of the CDL can be made more accessible to multiple audiences. Refinements of the crosswalks will certainly be informed by and deployed in the context of this IU-CDL collaboration.

In addition to the crosswalks per se, we will illustrate the use of crosswalks in the context of some popular desktop tools. A possible example will be deploying the crosswalks to import materials into Microsoft Office 11.

## **2. Exploratory Architecture and Prototypes for the Deployment of Crosswalks**

*We will design an architecture and software prototypes that build upon the use of direct crosswalks between formats.*

Although crosswalks are a good first step towards semantic interoperability, there are some significant limitations of this approach:

- Crosswalks share with all other formal automated computation techniques the inability to translate concepts from one domain to another when the concepts are not semantically equivalent in each domain.
- Because each crosswalk is hand-crafted, this approach does not scale well in handling large number of formats (If  $N$  is the number of formats,  $N*(N-1)$  crosswalks would be required to enable interconversion between any two formats.)
- The crosswalks do not handle translating pieces of documents. That is, these crosswalks do not provide for the disaggregation of documents.

To address some of these limitations, we will create software prototypes aimed at leveraging the crosswalks to the full extent. For example, a software prototype will automatically recognize the type of a given input document and enable users to choose from a menu of relevant applicable crosswalks. In the absence of a direct crosswalk, this program will use multi-step crosswalks to translate between any two formats that are translatable through a number of intermediate crosswalks (e.g., METS to RSS crosswalk through crosswalking METS to IMS-CP first and then IMS-CP to RSS). Certainly, a direct crosswalk between any two formats is preferable to the use of multi-step crosswalks. Since we will have direct crosswalks between the specifications discussed here, multi-step crosswalks will not be needed to handle such translations. However, as we consider other formats, it will become impractical to generate direct crosswalks for every combination, leading us to turn to multi-step crosswalks as one solution.

## **3. Relating Crosswalks to Other Approaches to Semantic Interoperability**

*We will write several versions of a white paper that will summarize our crosswalk investigations and survey promising approaches to enhancing semantic interoperability, looking for specific practical short-term applications of these approaches. We will also describe how crosswalks are related to these other techniques. We will publish the final version of the white paper in an appropriate venue.*

The lack of complete semantic interoperability will remain a problem for the foreseeable future. Although crosswalks are a practical approach to improving interoperability among a small number of specific domains, crosswalks are by no means the only way to address the problem. Two examples of other, more ambitious and long-term efforts to enable semantic interoperability among large numbers of arbitrary domains are:

- SIMILE ("Semantic Interoperability of Metadata and Information in unLike Environments") -- a recently funded \$4 million three-year collaboration among the MIT

Library (DSpace), the MIT CS department, and the W3C to leverage the connection among libraries, the semantic web, and personal information management. [2]

- The HARMONY project, a three-year project that investigated "a conceptual model for interoperability among community-specific metadata vocabularies." [1]

## **Role of Collaboration in the Proposed Work**

*We will assemble an Advisory Board of knowledgeable representatives (domain experts) from the library and educational technology communities as well as others who have been working at enhancing semantic interoperability among these domains. We seek advisors who are deeply familiar with the specifications as well as those experienced with production services that are or may use the specifications.*

The quality of our work on the crosswalks will be dramatically improved by working in dialog with such an Advisory Board. The Advisory Board will help guide the project so that it addresses the most salient issues and can therefore be of long-term value to the educational technology and library worlds. The Board will help ensure that the technical work is properly field-tested, and that this work on semantic interoperability is appropriately disseminated to various represented communities.

We will solicit the input of the Board about the general direction of the project at the beginning of our work. We plan to make well-defined, specific, and limited requests of the Board centered around the two versions of the crosswalks and the white papers. That is, we will ask the Board for feedback on the white papers and a subset of the Board to field-test the crosswalks. Most of our communication with Board members will be conducted via email and an internet-based discussion forum/website that the IU will host, as well as by phone. We will synthesize the feedback that we receive to shape the writing of the white papers and the crosswalks.

Because our communication with the Board may need to be supplemented by forums involving the entire Board or a significant number of Board members, we have budgeted for three videoconferences. Strategic times for staging the conference are 1) early on in the project (around the delivery of version 1 of the white paper and crosswalks) to allow for a group discussion on how the work should proceed for the next six months and 2) at the end of the project to discuss the future of this work beyond the project. We have also budgeted for travel to conferences to present the work of the project and for visits to specific members of the Board. Some of these conferences may be opportunities to gather small groups of Board members.

In addition to the formal feedback via the Web and email, we anticipate less formal ongoing collaborations with a number of the Board members, especially those who work on the Berkeley campus or in the San Francisco Bay Area with whom we already have working relationships. We are still finalizing the list of people we expect to invite.

*We would like to present this work at appropriate meetings of the library, educational technology, and internet technology communities.*

Possible meetings include the Digital Library Federation Spring Forum (May 2003) and Fall Forum (November 2003), the IMS Open Technical Forum and Library SIG (Summer or Fall 2003) and the Educause Annual Conference (November 2003). Raymond Yee

will be speaking at the O'Reilly Emerging Technology Conference in April 2003 on "University, Library, and Museum Content Meets XML, Web Services, and P2P" in an effort to spur conversation between the web technologists and the cultural heritage and educational communities.

## Anticipated Steps Beyond This Project

We anticipate that this project will lay a solid foundation for the practical deployment of semantic interoperability at UC Berkeley, while providing firmly grounded guidance to others in the educational and technology communities wrestling with similar issues. Version 2 of the XSLT-based crosswalks will be sufficiently field-tested and refined to be of value to anyone desiring to convert between the specifications targeted by the crosswalks. Since crosswalks do not require much technical infrastructure, the XSLT and related documentation will simply need to be archived by the IU (or other interested parties) to be available to others. However, for greater long-term usefulness of this project's work, we will encourage some of the organizations that produce interoperability specifications (for instance, the Digital Library Federation, the Library of Congress and IMS) to incorporate the crosswalks into their best practices around semantic interoperability with other specifications. Perhaps this work will enable better harmonization of future specifications developed in these communities, obviating future crosswalk work.

Since the IU will be incorporating the crosswalks in at least the early versions of the Scholar's Box, the IU will continue hosting, documenting, and refining the crosswalks. In the long term, as other approaches to semantic interoperability come to fruition and as the IU develops the Scholar's Box -- which will be designed to handle a broad range of materials from multiple repositories -- we may find that the crosswalks will be subsumed into or rendered obsolete by more general approaches that will then become timely to implement. In the short term, however, especially for those who have no interest in adopting larger frameworks, direct crosswalks will be important enablers of interoperability.

At the end of the project, we imagine that there will be promising and fruitful next steps to take. The specifics of these steps will, of course, depend on the findings of this work, the status of the Scholar's Box project, and the state of specifications development in the library and educational technology worlds. However, we anticipate that there will be interest in continuing the conversations between the library and educational technology communities that this project will catalyze and in institutionalizing our combined efforts. Should that be the case, UC Berkeley and the IU specifically will be interested in structuring such an effort and working with the Andrew W. Mellon Foundation in organizing such a project.

## Timeline

Overall timeline: March 2003- January 2004 (11 months)

Milestone	Date
V. 1 of crosswalks and documentation	March 31, 2003

Assemble Advisory Board	April 15, 2003
V. 1 of white paper on architecture of crosswalk and relationship of crosswalk to other semantic interoperability work	May 1, 2003
Presentation at DLF Spring Forum (Possible)	May 14-16, 2003
V. 2 of the white paper on architecture of crosswalk and relationship of crosswalk to other semantic interoperability work, to reflect prototype work and advisors' feedback	August 2003
Attendance at IMS Technical Forum	Summer or Fall 2003
V. 2 of crosswalks and documentation	November 2003
Presentation at DLF Fall Forum or related conference	November/December 2003
Final version of white paper	January 2004

## References

- [1] Harmony Main Page. <http://www.metadata.net/harmony/>
- [2] SIMILE Project. <http://web.mit.edu/simile/www/>
- [3] Heflin, J. and Hendler, J., Semantic Interoperability on the Web. in *Proceedings of Extreme Markup Languages 2000*, (2000), Graphic Communications Association, 111-120.
- [4] Hunter, J. MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains. *Journal of Digital Information*, 1 (8). <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/>
- [5] Moen, W.E., Mapping the interoperability landscape for networked information retrieval. in *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, (Roanoke, VA, 2001), The Association for Computing Machinery, 50-52. <http://www.unt.edu/wmoen/publications/MapInteropJCDLFinal.pdf>