Multi-Scale Science: Supporting Emerging Practice with Semantically Derived Provenance

James D. Myers,[1] Carmen Pancerella,[2] Carina Lansing,[1] Karen L. Schuchardt,[1] and Brett Didier[1]

[1]Pacific Northwest National Laboratory, Richland, WA 99352
[2]Sandia National Laboratories, Livermore, CA 94551-0969
jim.myers@pnl.gov, carmen@sandia.gov, carina.lansing@pnl.gov, karen.schuchardt@pnl.gov, brett.didier@pnl.gov

Abstract

Scientific progress is becoming increasingly dependent on our ability to study phenomena at multiple scales and from multiple perspectives. The ability to recontextualize third-party data within the semantic and syntactic framework of a given research project is increasingly seen as a primary barrier in multi-scale science. Within the Collaboratory for Multi-Scale Chemical Science (CMCS) project, we are developing a general-purpose, informatics-based approach that emphasizes "on-demand" metadata creation, configurable data translations, and semantic mapping to support the rapidly increasing and continually evolving requirements for managing data, metadata, and data relationships in such projects. A concrete example of this approach is the design of the CMCS provenance subsystem. The concept of provenance varies across communities, and multiple independent applications contribute to and use provenance. In the CMCS project, we have developed generic tools for viewing provenance relationships and for using them to, for example, scope notifications and searches. These tools rely on a configurable concept of provenance defined in terms of other relationships. The result is a very flexible mechanism capable of tracking data provenance across many disciplines and supporting multiple uses of provenance information..

Introduction

The systems approach to science, which emphasizes the connections among phenomena studied at different scales and by different disciplines, is causing dramatic changes in how scientific results are communicated. While traditional paper publication is sufficient when results are relatively independent, system science promotes a model in which incomplete and preliminary results from one study are used to guide new research, which in turn, may feed information back to the earlier work. This new dynamic drives a shift towards online community data systems and a growing emphasis on documenting data with enough information for it to be used intelligently by others without additional inputs. The specific 'metadata' required to enable such use, and the format via which it is to be conveyed, varies across disciplines and is rapidly evolving as more sophisticated experiments emerge. Developing community systems to support multi-scale research thus presents significant challenges.

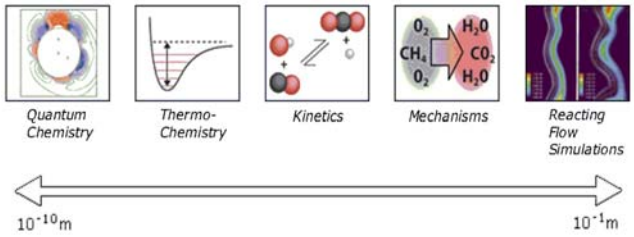

Figure 1. Multi-scale chemical science involves many physical phenomena and multiple research communities.

Data provenance (also called pedigree) provides a good example of the complexities involved. Provenance, in the most general terms, is metadata relating data to other data (versus, for example, just specifying its type or size). However, a wide range of possible definitions exists. Provenance can include the literature references where data were first reported; its history in terms of how it has been stored, curated, and transferred; the series of experimental procedures, computations, and/or database queries by which it was derived from other data; the sequence of ideas leading to an experiment; an experiment's relation to others as part of an overall project; its association with calibration and control data; etc. Which information is required, at what level of detail, and whether it is currently cost effective to record it, vary by type, purpose, discipline, and project [1]. A software approach that requires the meaning and format of provenance information to be standardized is unlikely to meet the needs of a multi-scale research community. We believe that acknowledging the real differences across such a community in the requirements for provenance information and the meaning ascribed to it, and designing community software to help manage the complexity, is a much more promising approach.

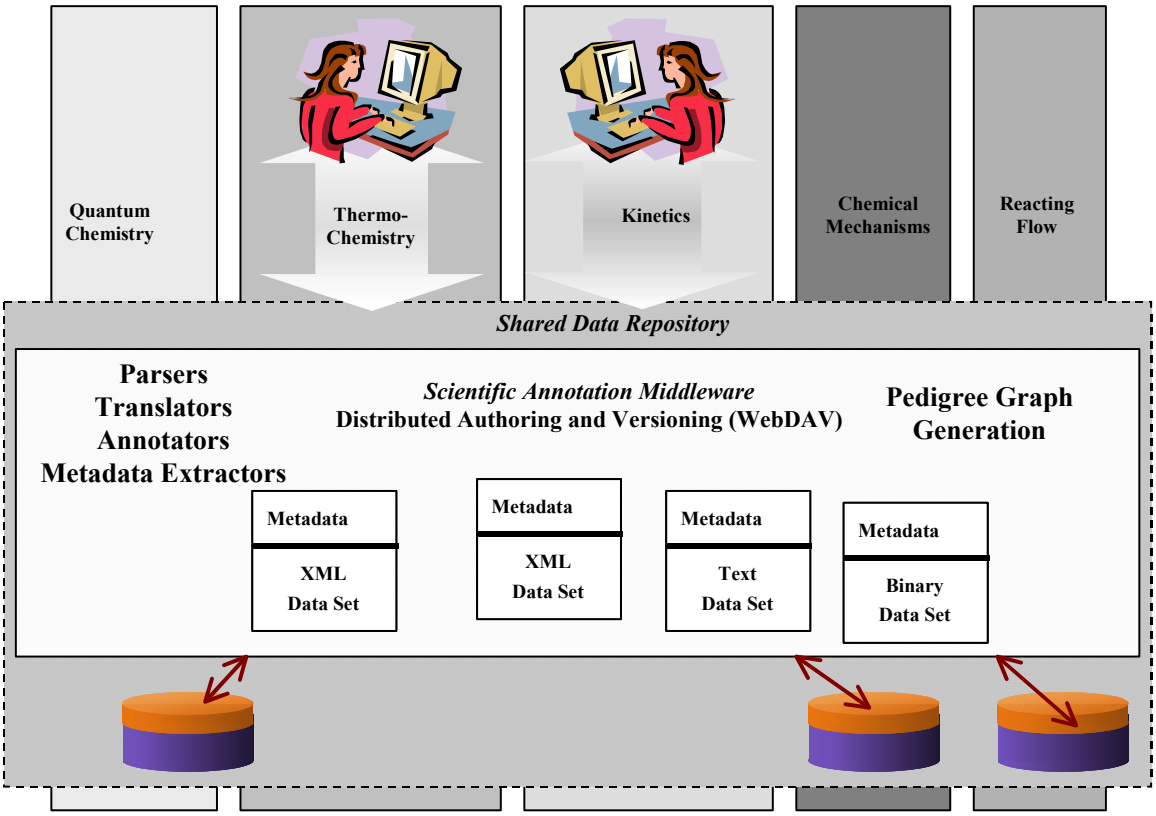

Figure 2. CMCS Infrastructure for Data/Metadata Management.

Collaboratory for Multi-Scale Chemical Science

The Collaboratory for Multi-Scale Chemical Science (CMCS) [2] is a multi-institution project funded by the U.S. Department of Energy to develop and pilot test an advanced collaborative community data system for chemical science. The collaboratory is intended as an open, public resource supporting a systems approach to chemical science including sub-disciplines ranging from quantum chemistry to device-scale chemical combustion (Figure 1).

CMCS includes tools for discovering, analyzing, and visualizing data as well as for coordinating group and community processes involving the data. The overall architecture of CMCS is shown in Figure 2. The main interface is the Multi-Scale Chemistry (MCS) portal. Within the portal, tools are available for exploring data collections, searching for chemical information about particular species, subscribing to receive email messages when new data appears, viewing metadata and data relationships, and visualizing data. Additional tools, such as electronic notebooks, discussion forums, task lists, and calendars, are provided for group coordination. The portal is built using the CompreHensive collaborativE Framework (CHEF) [3], which itself leverages the Apache Jetspeed portal framework.

The general approach used in CMCS to manage data/metadata is guided by the following design decisions [4,5]:

- Scientists should not be forced to adopt data standards to use the system. Rather, the system should support communities that do standardize while providing tools to map and translate information for those that do not wish to change the way that they do research or their data formats.
- CMCS should adopt the working definition that metadata is simply data that has meaning across domains and will be managed as WebDAV properties. Metadata is commonly defined as information 'about' data values and data sets. However, such a definition is very dependent on one's perspective and such differences of opinion, once encoded in software, are an endless source of barriers to cross-scale collaboration.
- CMCS will recommend an evolving core schema for scientific metadata within the environment. The core schema should be sufficient to support the data discovery, browsing, annotation, and visualization/ analysis capabilities of CMCS's standard tool set. As new tools are added to CMCS, we expect this core schema to be expanded to include properties to support these new capabilities, including, for example, third-party annotation and scientific peer review.
- CMCS users should be able to experiment with more detailed metadata and additional uses for metadata.

Underlying the portal, CMCS has developed an advanced data repository that supports these decisions and automates many aspects of data discovery, translation, and data provenance tracking. CMCS employs the Scientific

Annotation Middleware (SAM) as a means to provide federated data/metadata access, extensible metadata annotation, and transformations of data and metadata [6]. SAM is based on Jakarta Slide an open-source Java content repository that implements the Web Distributed Authoring and Versioning (WebDAV) protocol extensions to HTTP [7]. As a WebDAV capable server, SAM accepts arbitrary typed data and arbitrary text/XML metadata [8]. SAM can be configured to map this content to underlying databases or Grid-based data repositories.

Defining Metadata and Relationships

Data can be submitted to CMCS via the WebDAV protocol or manually via the MCS portal. Metadata can be associated with the data in several ways. Applications can submit metadata directly to the repository via WebDAV. The MCS portal provides a form-based interface for manual annotation. SAM can also invoke registered transformations to automatically generate user-defined metadata. XSLT scripts can be used to extract metadata from within XML-formatted files and populate desired WebDAV properties. The Binary Format Description (BFD) language [9] provides a similar capability to extract properties from binary or ASCII formatted data. SAM can also be configured to call external web services for more complex operations.

In addition to simply extracting metadata, transformations can also map between schema used within the document to the CMCS core schema or to those used by other applications or in other domains. CMCS has developed a variety of metadata generators for chemical data types from a variety of scales. These transformers generate basic metadata such as "creator" as defined in the Dublin Core Element Set [10] as well as data-type-specific information, such as the chemical reaction to which the data refer [11]. In essence, these translators define the semantic mapping between the source application and the CMCS tools and/or applications in other domains.

The decision of whether to use metadata generators or to populate metadata programmatically is essentially left to the application developer. Both methods have been used in CMCS. For data manually uploaded to CMCS via the portal and applications using WebDAV within a file metaphor (using only the WebDAV PUT/GET methods or using a file system driver such as WebDrive [12]), metadata generators have been preferred. An example of the latter is the Active Thermochemical Tables (ATcT) application [13], which has been integrated with CMCS via a portlet and web service. The ATcT web service, which reads and writes several types of ATcT data files to/from the CMCS repository via WebDAV, uses registered XSLT scripts to generate CMCS-related properties. This approach allows ATcT to maintain file semantics and retain a degree of independence from CMCS that would simplify use of ATcT in other contexts. Other CMCS communities, working with large numbers of manually created files have chosen to use metadata generators for similar reasons.

Conversely, some community chemistry applications, such as the Extensible Computational Chemistry Environment (Ecce) [14], and the domain-independent electronic laboratory notebook (ELN) [15], both of which have an internal distinction between data and metadata, populate CMCS metadata directly. Ecce, a WebDAV-aware problem solving environment for molecular-scale computational chemistry, directly generates properties in the CMCS schema related to the creator, chemical species, and relationships among the large number of files produced during a computational chemistry calculation. Similarly, the ELN records information such as author names, data types, chapter/page/note relationships, and digital signatures directly as WebDAV properties compatible with CMCS schema and conventions (discussed next).

CMCS uses several conventions to support interpreting WebDAV properties as semantic relationships. The most basic convention is an implicit assumption that properties are "about" the resource with which they are associated. CMCS also uses a convention that WebDAV property names should be phrased as verbs when possible or have an obvious verb form. For example, "issanctionedby" is considered a verb, while "creator" implies a "created by" verb. Property values are interpreted as the object in relationships. Currently, CMCS assumes that property values are literal unless they contain an XLink [16] in which the resource at the URL defined in the XLink is treated as the object. Since WebDAV property names must be unique, CMCS uses the convention of interpreting multiple items within an RDF container element to imply multiple relationships that share a common subject (the resource) and verb (the property name).

These conventions are aimed at minimizing users' exposure to semantic languages such as RDF while still allowing some level of human and machine interpretation of metadata. To incorporate semantic tools within its infrastructure, CMCS plans to use SAM to dynamically generate an RDF metadata document based on a resource's WebDAV properties. As an initial experiment in this direction, we have implemented a dynamic WebDAV property for which the value is the "literal" conversion of the other properties to XML-encoded RDF (no further mapping to particular RDF schema or other verification that CMCS conventions were followed).

Viewing Provenance in CMCS

Metadata defined in the CMCS repository can be accessed directly via WebDAV through generic browsers (such as DAVExplorer) or custom WebDAV-aware applications. They can also be viewed as data relationships directly within the CMCS portal using the "Pedigree Browser" portlet. A screenshot of this browser can be seen in Figure 3. The portlet uses the conventions discussed previously, a map from property names (and qualifying namespaces) to

more readable labels, and an automated XSLT transform to generate an HTML display of the current resource's metadata. Properties that refer to other resources are rendered as HTML links. Clicking a link changes the focus of the Pedigree Browser to the new resource, thus enabling users to manually traverse the pedigree tree. The latest version of the Pedigree Browser also provides, in the left pane, a mechanism for showing or hiding properties in different XML namespaces. The right pane then displays the formatted values for the selected properties.

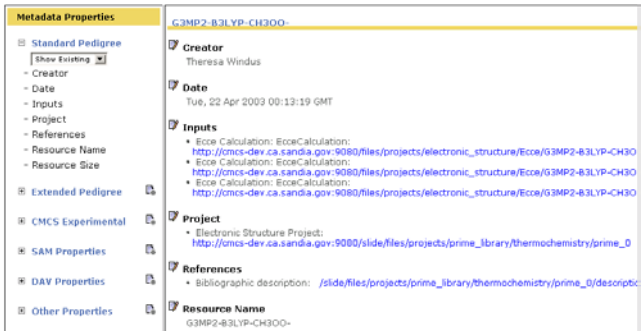

Figure 3. Central panes of the CMCS Pedigree Browser showing the metadata and relationships of the selected data set.

While the Pedigree Browser is a powerful tool, even our initial pilot testing showed the limitations of a browsing metaphor for understanding long or complex relationship chains. Further, we discovered minimal consensus between groups over which metadata should be standard and what level of detail is required in defining semantics, suggesting that a general, coarse-grained mechanism for pruning the properties shown would not be sufficient in the longer term.

As a result, we developed the "Pedigree Graph" portlet, shown in Figure 4, and took its development as an opportunity to investigate a more flexible mechanism for defining which relationships should be shown and how mapping between schema, and between levels of detail, could be accomplished. The portlet itself dynamically generates a Scalable Vector Graphics (SVG), JPEG, or GIF image based on a description of the relationships described in the Graphical eXchange Language (GXL) format. The GXL input is dynamically generated by SAM based on a definition of provenance in terms of other relationships. (An RDF format output is also available.) The definition is currently captured in an XML document that describes which properties are "a type of" provenance and the overall depth of relationship chains to report.

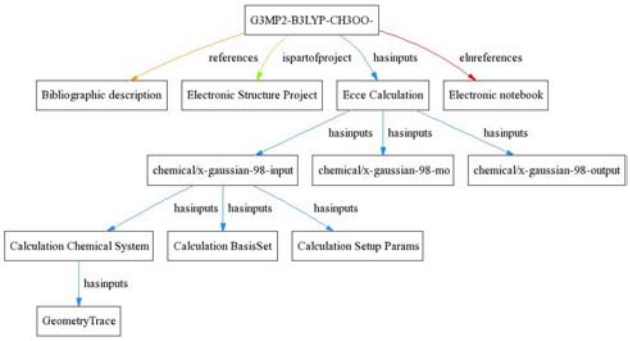

Figure 4. CMCS "Pedigree Graph" portlet showing provenance relationships between resources (color coded by original relationship type).

As currently configured for the pilot CMCS server, the graph reports relationships (to a depth of five) of the following types (where namespaces have been abbreviated for clarity):

- cmcs:hasinputs—links a resource to one or more inputs
- cmcs:hasoutputs— links a resource to derived data
- cmcs:ispartofproject—links data or a collection of data to a project description
- dcterms:references—links a resource to other resources that it references (a Dublin Core metadata element)
- dcterms:isreferencedby—links a resource to resources that reference it (a Dublin Core metadata element)
- sam:hastranslations—shows a list of translations that can be created from a resource using SAM
- eln:children—shows children relationships in an electronic notebook
- eln:elnreferences—shows reference relationships to an electronic notebook

This definition was chosen to provide a useful visual summary of scientifically relevant dependencies (i.e., changes to related nodes could affect the validity/interpretation of the current resource). Similarly, the depth was set at five to limit visual clutter. We anticipate that users will need to refocus the browser on "leaf" nodes to follow longer provenance chains.

With the graphical view, CMCS users can see relationships generated by multiple independent applications and discern connections that would not be evident within a single application. For example, it would be possible to see that two calculations being documented in a group notebook may be using different versions of reference data as inputs. Such a graphical view of provenance relationships across scientific scales is a powerful capability; currently researchers spend considerable time assembling such information. CMCS already contains data with provenance trails that extend from kinetic mechanisms to reaction rates and the thermochemistry of individual chemical species and to the original quantum calculation from which the thermochemistry is derived, involving both data and

literature references (Figure 4 shows part of such a multi-scale provenance chain).

As researchers become familiar with using the Pedigree Browser and as additional applications begin to contribute new types of metadata, we expect that the definition of provenance used by the browser will evolve and eventually become configurable by individuals and groups and, perhaps, specialized by purpose.

Using Provenance Programmatically

SAM can currently generate either GXL or RDF output describing the provenance information and it is being extended to accept the definition of provenance to be used on a per-request basis (versus a per-server setting). With these capabilities, users and tools will be able to maintain their own definitions of provenance and request information from the CMCS repository consistent with their conceptual model.

Thus, a workflow tool could request provenance information confined to the detailed processing history of a data set while a notebook might define provenance in terms of project and task relationships and use it to dynamically structure a project notebook with the same organization. The development of external agents, which could reason across the RDF triples they understand, would also be possible. Such agents might then derive new inferred relationships and automatically create new relationship links.

Another capability in development is support for the use of provenance definitions in scoping search and notification operations. Currently, CMCS search queries and email notification services are scoped implicitly via the WebDAV collection hierarchy (i.e., the operations apply recursively to a branch of the CMCS URL namespace). We are currently investigating ways to allow these services to use an explicit provenance definition instead. This enhancement would allow researchers to request an email notification if any of the data used as input in their current work is updated, or to search just within experiments collected into their current notebook.

While many of the use cases we have considered involve simple, many-to-one, is-a-type-of mappings, we do anticipate a need for more general capabilities to infer new relationships from existing ones. Specifically, inferences about the existence of inverse and transitive relationships will clearly be useful. For example, a citation-reporting tool might infer a provenance relationship between two published data sets based on the existence of a chain of "hasinputs" relationships, which were in turn inferred based on directly recorded "hasoutputs" relationships.

Conclusion

As part of its overall infrastructure, CMCS is developing a very flexible model for managing scientific provenance from a systems, and systems-science perspective. The model assumes multiple, independent actors (researchers, agents, applications) that, while sharing some concept of data provenance, never-the-less differ in the details of their perspective and the level-of-sophistication of their conceptualization. The CMCS data repository includes configurable services for extracting metadata from arbitrary data formats as well as emerging capabilities for dynamically inferring additional relationships. The CMCS portal includes multiple tools that document data relationships and general capabilities for browsing and visualizing provenance.

These basic capabilities are currently in pilot use by international groups of chemists engaged in research to assemble a new generation of "living" chemical reference information. Their initial requirements and ongoing feedback are being used to guide and prioritize our development of richer descriptions of provenance and the adoption of the semantic technologies required to process them.

We believe that the approaches and technology that CMCS is piloting will not only allow increased collaboration and coordination across scientific disciplines and chemistry scales, but will be a critical enabler of new, potentially revolutionary, approaches to scientific inquiry.

Acknowledgements

This work was supported as part of the Collaboratory for Multi-Scale Chemical Science (CMCS) project within the National Collaboratories Program sponsored by the U.S. Department of Energy's Office of Mathematical, Information, and Computational Sciences. Sandia National Laboratories is operated by the Sandia Corporation under Contract No. DE-AC04-94-AL85000 with the U.S. Department of Energy. Pacific North Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract No. DE-AC06-76RLO 1830.

The authors wish to acknowledge CMCS team members from the following institutions for their contributions to the overall CMCS framework and concept: Sandia National Laboratories, Pacific Northwest National Laboratory, Argonne National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, National Institute of Standards and Technology, Massachusetts Institute of Technology, and University of California, Berkeley.

References

[1] Buneman, P., Khanna, S., and Tan, W.C. 2001. "Why and Where: A Characterization of Data Provenance." Proceedings of the International Conference on Database Theory (ICDT), 8th International Conference, London, UK, January 4-6, 2001.

[2] The Collaboratory for Multi-Scale Chemical Science Website, http://cmcs.org/

[3] CHEF Collaborative Portal Framework, http://www.chefproject.org/
[4] Pancerella, C., Myers, J., and Rahn, L. 2002. "Data Provenance in the Collaboratory for Multi-scale Chemical Science (CMCS)", Workshop on Data Derivation and Provenance, October 17-18, 2002, Chicago, Illinois. Website: http://people.cs.uchicago.edu/~yongzh/position_papers.html
[5] Myers, J. 2002. "Design Constraints for Scientific Annotation Systems." Workshop on Data Derivation and Provenance, October 17-18, 2002, Chicago, Illinois, Website: http://people.cs.uchicago.edu/~yongzh/position_papers.html
[6] Myers, J.D., Chappell, A., Elder M., Geist A., and Schwidder, J. May/June 2003. "Re-Integrating the Research Record." IEEE Computing in Science and Engineering, 5(3): 44-50. Available at http://www.scidac.org/SAM/
[7] Jakarta Slide Java Content Management System Website, http://jakarta.apache.org/slide/index.html
[8] Web Digital Authoring and Versioning (WebDAV) Resources Community Website: http://www.webdav.org/
[9] http://collaboratory.pnl.gov/docs/collab/sam/bfd/
[10] Dublin Core Metadata Element Set, Version 1.1: Reference Description, 2003, from Dublin Core Metadata Initiative Website: http://www.dublincore.org/documents/dces/
[11] Pancerella, C., Myers, J., Allison, T., Amin, K., Bittner, S., Didier, B., Frenklach, M., Green, W.Jr., Ho, Y-L., Hewson,, J., Koegler, W., Lansing, C., Leahy, D., Lee, M., McCoy, R., Minkoff, M., Nijsure, S., von Laszewski, G., Montoya, D., Pinzon, R., Pitz, W., Rahn, L., Ruscic, B,. Schuchardt, K., Stephan, E., Wagner, A., Wang, B., Windus, T., Xu, L., Yang C, 2003. "Metadata in the Collaboratory for Multi-Scale Chemical Science." Proceedings of the 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research and Applications (DC 2003), September 28-October 2, 2003, Seattle, Washington.
[12] Webdrive FTP, WebDAV, and Internet File Manager, South River Technologies, Website: http://www.webdrive.com/products/webdrive/index.html
[13] von Laszewski, G., Ruscic, B., Wagstrom, P., Krishnan, S., Amin, K., Nijsure, S., Pinzon, R., Morton, M.L., Bittner, S., Minkoff, M., Wagner, A., and Hewson, J.C. 2002. "A Grid Service Based Active Thermochemical Table Framework." Third International Workshop on Grid Computing, Lecture Notes in Computer Science, November 18, 2002, Baltimore, Maryland.
[14] Schuchardt, K. L., Myers, J. D., and Stephan, E. G. 2002. "A Web-based Data Architecture for Problem Solving Environments: Application of Distributed Authoring and Versioning to the Extensible Computational Chemistry Environment." Cluster Computing, 5: 287-296.
[15] Myers, J., Mendoza, E., and Hoopes, B. 2001. "A Collaborative Electronic Notebook." Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2001), August 13-16, 2001 Honolulu, Hawaii.
[16] XML Linking Language (XLink) Version 1.0, June 2001, from World Wide Web Consortium Website: http://www.w3.org/TR/xlink/