NSF / NSDL Workshop:
Scientific Markup Languages
Workshop Report

FINAL REPORT

THE NATIONAL SCIENCE DIGITAL LIBRARY

Hosted by the National Science Foundation
Arlington, Virginia
June 14-15, 2004

Sponsored by the National Science Foundation
under grant no. DUE-0333520

Report prepared by:

Laura M. Bartolo, Kent State University
Timothy W. Cole, University of Illinois at Urbana-Champaign
Sarah Giersch, Association of Research Libraries
Michael Wright, UCAR – DLESE Program Center

Table of Contents

Workshop Organizing Committee

- Laura Bartolo, Kent State University, Planning Committee Chair, Materials Digital Library
- Howard Burrows, Autonomous Undersea Institute, Policy Committee Chair, NSDL
- Stuart Chalk, University of North Florida, Analytical Sciences Digital Library
- Tim Cole, University of Illinois Urbana Champagne, Planning Committee Co-Chair, Second Generation Digital Mathematics Resources
- Ben Domenico, UCAR – Unidata, THREDDS Second Generation
- Sam Dooley, Integre Technical Publishing Co., Inc., techexplorer and MathDL
- Sarah Giersch, Association of Research Libraries, iLumina Digital Library
- John Saylor, Cornell University, NSDL Core Integration
- Mike Wright, UCAR – DLESE Program Center, Digital Library for Earth Science Education

Acknowledgements

The workshop organizers would like to thank the NSF Directorate for Education and Human Resources (EHR) and the Division of Undergraduate Education (DUE) for supporting the workshop, with special thanks to Dr. Lee L. Zia, Program Director for the National Science Digital Library (NSDL), for support and encouragement. Thanks also to the workshop attendees for furthering and expanding scientific markup languages by providing perspectives about markup languages through written statements, participating in the workshop, and commenting on this report.

Workshop Participants

Name	Company / Institution
Laura Bartolo	Kent State University
Henry Bass	MatWeb
Judy Brown	Academic ADL Co-Laboratory
David Burggraf	Galdos Systems Inc.
Howard Burrows	AUSI
John Caron	UCAR - Unidata
Tim Cole	University of Illinois at Urbana-Champaign
Donald DeLand	Integre Technical Publishing
Philip Dodds	Advanced Distributed Learning
Ben Domenico	UCAR - Unidata
Sam Dooley	Integre Technical Publishing
Mark Doyle	The American Physical Society
Dave Dubin	University of Illinois Urbana-Champagne
Tim Finin	University of Maryland, Baltimore County
Sarah Giersch	Association of Research Libraries
Nancy Gough	AAAS
Scott Henry	American Society of Metals
Ted Hodapp	National Science Foundation
Andy Hunt	Wolfram, Inc.
Tim Ingoldsby	American Institute of Physics
Patrick Ion	Mathematical Reviews
Brett Johanson	Boeing
Bob Kelly	The American Physical Society
Gary W. Kramer	NIST
David Martinsen	American Chemical Society
Dave McArthur	National Science Foundation
Robert Miner	Design Science, Inc.
Peter Murray-Rust	Cambridge University
Stefano Nativi	University of Florence
John Phipps	Cornell University
Jim Pillars	Boeing
Dominik Poetz	NIST
Adam Powell	MIT
Rahul Ramachandran	University of Alabama, Huntsville
Rob Raskin	NASA Jet Propulsion Laboratory
Hal Richtol	National Science Foundation
John Rumble	Information International Associates
Daniel Tofan	SUNY Stony Brook
Richard Ullman	NASA
James Warren	NIST
Dick Wertz	Foundation for Earth Science
Mike Wright	UCAR - DLESE Program Center
Lee Zia	National Science Foundation

Executive Summary

This report summarizes a workshop on scientific markup languages (MLs), sponsored by the National Science Foundation, June 14-15, 2004. The workshop goals were to assess and document scientific disciplines' work on markup languages and to begin to articulate a vision for the future evolution and implementation of markup languages in support of a cyberinfrastructure for research and education, with a particular focus on using markup languages in the context of the National Science Digital Library (NSDL).

The workshop opened with presentations that 1) provided a framework for the workshop discussions about scientific markup languages as they relate to the broader development of a knowledge infrastructure (e.g., the semantic web) and that 2) suggested that there is an ongoing tension between static data exchange standards and the dynamic nature of science, science research and scientific data.

Presentations on the current state of scientific MLs as used in four specific scientific domains (chemistry, earth sciences, materials sciences and mathematics) highlighted the idea that for MLs to move forward in a discipline, adoption and development must occur among communities of scientists, publishers and vendors, and end-users simultaneously.

Cross-domain discussions around topics (Education, Markup Languages, Publishers / Professional Societies, and Database / tool developers) identified several cross-cutting themes and recommendations:

Theme A: Vision. Motivating the development of markup languages that are built on XML is the belief that by providing a means to exchange information, or data, in a structured form that colleagues across scientific domains can read, understand and use, scientific research and discovery can be moved forward. Through common interoperability mechanisms, NSDL supports the exchange of information between the sciences and provides a framework for markup languages to be extended even further as they are tested and applied in science education settings.
Recommendation 1: NSDL should play a central role in organizing cross-domain work on markup languages
Recommendation 2: Continue support for cross-domain community interaction

Theme B: Demonstrating the value of markup languages. Despite the potential to benefit several science and research applications, markup languages' value in those contexts remains unproven. Their broadest implementation to date occurs in processes that are virtually invisible to most users.
Recommendation 3: Support assessing the potential benefits of markup languages

Theme C: Creating & disseminating the pre-requisite tools. Better tools, both technically and in the form of broader, more robust ontologies, would facilitate and speed the adoption of scientific markup languages.

Recommendation 4: Conduct an environmental scan of scientific markup language tols and ontologies
Recommendation 5: Support applied research to produce needed tools and ontologies

Theme D: Mediation of markup languages. "Mediation" covers the concept of tools and services that provide a translation interface between representations in different markup languages, or that provide access to information in a single markup language to a wide variety of users.
Recommendation 6: Support research on mediation services and tools between markup languages
Recommendation 7: Support research on services and tools that mediate between markup languages and end users in education
Recommendation 8: Work with appropriate organizations to encourage to conclusion the development of UnitsML.

Theme E: Identifying challenges to maturation of markup languages. There are cultural and market-related challenges to sustaining an attenuated consensus-building process around scientific markup languages.
Recommendation 9: Support the next stage of scientific markup language standardization and implementation
Recommendation 10: Fund targeted needs assessments to identify audience(s) for scientific markup languages

Specific actions items from the workshop include:

- Continuing the workshop listserv to support ongoing cross-domain discussions
- Develop a registry of scientific markup languages
- Plan a follow-on workshop in 2005

Note about the Preliminary Draft
A draft version of the Workshop Report was circulated among workshop attendees for comments.

NSDL Annual Meeting 2004 Follow-Up
A panel discussion entitled "Use of Scientific Markup Languages in the NSDL" was held at the 2004 NSDL Annual Meeting. The panel abstract and presentations are available at: http://nsdl.comm.nsdl.org/meeting/schedules/schedule.php?proposal_id=2639&nsdl_annual_meeting=effeccc1e40c90e201f8ced3720808a9

Introduction

This report summarizes a workshop on scientific markup languages (MLs), sponsored by the National Science Foundation, June 14-15, 2004. The workshop brought together forty-three higher education, publishing and software, and government representatives from the disciplines of biology, chemistry, earth sciences, mathematics, materials sciences and physics. The workshop goals were to assess and document scientific disciplines' work on markup languages and to begin to articulate a vision for the future evolution and implementation of markup languages in support of a cyberinfrastructure for research and education, with a particular focus on using markup languages in the context of the National Science Digital Library (NSDL).

Digital libraries are catalysts for new knowledge that provide content and tools for successive researchers, such as students, educators, or scientists, to build upon the findings of prior users, and that maintain existing information archives which can inform new discoveries. Advances toward the Semantic Web and a distributed cyberinfrastructure promise revolutionary changes for the way science and engineering will be conducted in education, research, and industry. A critical underpinning of both the Semantic Web and cyberinfrastructure is the use of scientific markup languages. As digital libraries become fully integrated into the work and research of students, educators, and scientists, they can play a pivotal role in the development of the Semantic Web and cyberinfrastructure.

About NSDL

The National Science Digital Library (NSDL) Program was launched by the National Science Foundation in 2000 to establish an online library of exemplary resources for science, technology, engineering, and mathematics (STEM) education and research. NSDL provides organized access to collections and services from resource contributors that represent the best of public and private institutions including universities, museums, commercial publishers, government agencies, and professional societies. NSDL supports teaching and learning at all levels with materials ranging from journal articles and lesson plans to interactive animations and from real-time data sets to technology-based tools. With a mission to support national improvements in STEM education and an emphasis on innovation, NSDL began in the fall of 2000 to build the technical infrastructure of the Library, coordinate access to resources from a wide range of providers, and build relationships with key stakeholders in the research and education communities. Access to aggregated NSDL collections and services began with the launch of the *NSDL.org* Web site in December 2002. Since NSDL was established, six NSF/NSDL-sponsored workshops have addressed pivotal issues that impact the communities represented in NSDL.[1]

To date, development and evolution of systems for encoding scientific data have largely taken place in isolation from one another, as well as from users. Disciplines developing MLs face common challenges, and MLs have progressed to the point where it is reasonable to address challenges and opportunities jointly. However, no organization that represents a broad spectrum of STEM MLs currently exists to bring the various languages together; NSDL is uniquely positioned to fill this role. NSDL is comprised of many digital libraries from various sectors and disciplines, such as biology, chemistry, earth sciences, mathematics, materials sciences, and physics, all of which are closely involved with the use of one or more scientific MLs. Building on discussions from the NSDL 2003 Annual Meeting, which identified several benefits for NSDL and for the scientific community for hosting a workshop on scientific MLs, this workshop was organized to bring together representatives from various stakeholder communities to begin to identify common opportunities and challenges for use of scientific markup languages in the science and education communities.

This workshop was the first step in an ongoing collaborative process that will facilitate faster, better, and more adaptable resolution of common issues, resulting in high-levels of interoperability between and among scientific markup languages and generating far-reaching future initiatives that support the integration of research and education. To that end, the workshop organizing committee identified representatives from a diverse range of science disciplines who have been active in developing markup languages. They participated by contributing statements outlining their discipline's needs or issues concerning markup languages and how markup languages could be used in NSDL to facilitate the integration of research and education. The diversity of workshop participants, including both those within and external to the NSDL community was key to insuring the quality and completeness of the breakout discussions. Interest in the workshop exceeded space available, and there are plans to continue discussions via a listserv and by convening a follow up workshop. The workshop agenda, details about breakout sessions, and URLs of presentations are included in the report's appendices and online at the workshop website.[1]

Making the Web Safe for Intelligent Agents

In his keynote presentation "Making the Web Safe for Agents," Tim Finin[2] addressed the issue of how to make the web machine-usable and why it should be done. He provided a framework for the workshop discussions about scientific markup languages as they relate to the broader development of a knowledge infrastructure that supports software agent applications.

> "The web has made people smarter. We need to understand
> how to use it to make machines smarter, too."

[1] Educational Publishers and NSDL (Oct 2002); Evaluating the Educational Impact of NSDL (Oct 2003); Exploring Business Options for NSDL (Nov 2003); Participant Interaction in Digital Libraries (Feb 2004); Scientific Markup Languages (June 2004); Developing a Web Analytics Strategy for NSDL (Aug 2004)

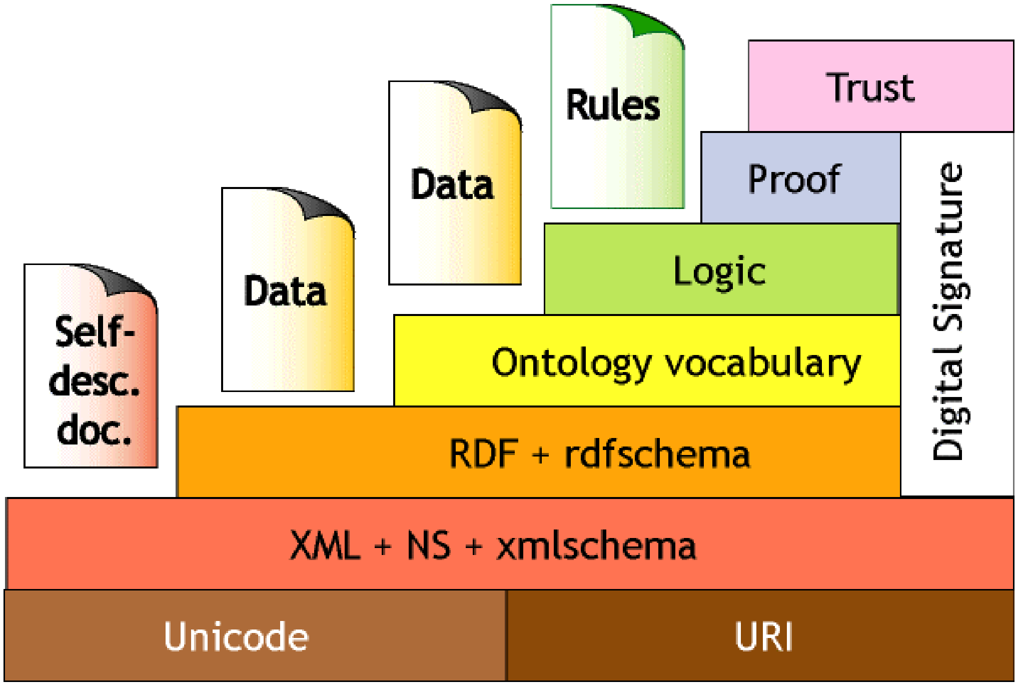
[1] http://scimarkuplang.comm.nsdl.org/

[2] Tim Finin is a Professor in the Department of Computer Science and Electrical Engineering at the University of Maryland, Baltimore County. He has over 30 years of experience in the applications of Artificial Intelligence to problems in information systems, intelligent interfaces and robotics, and is currently working on software agents. A link to his presentation, and others referenced in this report, are found in Appendix A.

- Michael I. Jordan, AAAI, July 2002.

Building on this quote, Finin noted there is a need to help software agents[3] benefit from the web. There is much discussion about building tools that find information on the web, that develop knowledge from the search process, and that help people understand information in the context of a problem. However, the current web does not support software agents by its inability to provide, among other things, machine usable and understandable interfaces. For the current web to be usable by software agents, several layers of languages, data and ontologies must be in place first. The first layer consists of XML markup languages, followed by a second layer of data over the web. This foundation allows machine-understandable structures to be built that will provide the basis for the future software agent applications. The next step is to develop a layer of knowledge representation. The semantic web (See Figure 1) is the beginning of that process.

Figure 1: Layers of the Semantic Web



The semantic web is the first serious attempt to provide semantics for XML sublanguages, and eventually it will provide mechanisms for people and machines (agents, programs, web services) to come together in all kinds of networked environments (e.g., wired, wireless, ad hoc, wearable, etc). A key component of the

[3] Software agents differ from conventional software in that they are long-lived, semi-autonomous, proactive, and adaptive.

semantic web is ontologies – theories of what exists. Information systems have adopted ontologies from philosophy and formalized them into specifications for use in applications (e.g., UML diagrams, data dictionaries, database schemas, conceptual schemas, and knowledge bases).

The current semantic web languages are RDF (Resource Description Framework: a language of simple subject-predicate-object triples building directed graphs) and OWL (Web Ontology Language: adding capabilities common to description logics such as cardinality). The choice of RDF or XML is not an either-or question; each has value. For example, the hierarchical tree nature of XML may be useful where applications may rely on hierarchy position. In this case, XML provides a simple syntax and structure. RDF can provide loose collections of relations that are easy to combine into one big set. The application of either will be greatly enhanced by the availability of tools to support markup languages, but current methods and tools miss the opportunity to do this. For example, currently, web content derived from databases only holds the content to be rendered as HTML; business chart tools don't explicitly save the structural relationships rendered in the chart view. There is a need for tools to be "semantic web aware."

There are a number of applications using RDF[4], RSS (RDF Site Summary - formerly called Rich Site Summary) and FOAF[5] (friend-of-a-friend) as examples of initial work building on the semantic web ideas. There are also several research issues to pursue (e.g., ontology alignment and mapping, learning ontologies, automating markup, extending OWL for rules and query languages, and integration with agents, web services and information retrieval).

In closing, Finin noted that it would take time to deliver on the intelligent agent paradigm either on the Internet or in a pervasive computing environment. The development of complex systems is an evolutionary process. He recommended that researchers and developers start with the simple and move toward the complex (e.g. from vocabularies to full ontology language theories), allow many ontologies to bloom, and support a diversity of ontologies since monocultures are unstable.

An Historical Perspective on Markup Languages

The primary objective of the workshop's first session was to provide context and a starting point for the day's work. The morning's plenary presentation[6] focused on placing scientific markup languages, considered as tools to an end, in the context of the broader

[4] RDF application in web standards: CC/PP (Composite Capabilities/Preference Profiles); P3P (Platform for Privacy Preferences Project); RSS (RDF Site Summary); RDF Calendar (~ iCalendar in RDF).
RDF application in other systems: Netscape's Mozilla web browser; Open directory (http://dmoz.org/); Adobe products via XMP (eXtensible Metadata Platform); Web communities: LiveJournal, Ecademy, and Cocolog

[5] Applications of FOAF: http://rdfweb.org/topic/ApplicationIdeas; also: http://www.foaf-project.org/

[6] John Rumble, Information International Associates, gave the workshop opening plenary presentation entitled, "The Dynamics of Data Standards."

body of scientific standards which facilitate and in some cases enable the more effective reporting, dissemination, and use of scientific data and analyses. In particular, John Rumble suggested that there is an ongoing tension between static data exchange standards and the dynamic nature of science, science research and scientific data. XML-based markup languages, because they can be defined to include a measure of flexibility and extensibility, have the potential to help bridge that dynamic tension. To achieve that potential, developers of scientific markup languages should examine how successful scientific standards historically have tended to evolve and attain prominence, the motivations of standards users and adopters, the nature of the data exchange tasks being attempted through the use of markup languages, and the lessons to be learned from studies of language evolution more generically (e.g., the study of how human languages have evolved). In his presentation Rumble discussed all of these themes in turn. To give a flavor of his presentation and provide context for this report, a few of his most salient points made are paraphrased here. (You can also see them echoed in the discussions and recommendations that came out of the workshop.)

Technical standards in the sciences have a well-recognized life cycle. Simplified, most technical standards begin when an individual practitioner or a community collectively recognizes and articulates the need for a consensus way to perform a common task. To address that statement of need, a group of expert practitioners convene. Technical solutions are proposed, differences hashed out, and eventually a consensus emerges. The standard is published and adopted by users. If the adoption base is sufficient and the business model sufficiently robust, enough support is generated to maintain the standard, which is then over time refined, republished, and re-implemented by users as necessary.

Adoption and broad use is crucial to achieving success in the development of a standard. And adoption in turn requires that users be motivated to adopt. In considering markup languages as technical data exchange standards, the fact that they will enhance interoperability and facilitate long-term archiving is arguably not enough for most users, especially for scientists authoring and ultimately using data conveyed in marked up form. Simply achieving interoperability and consensus with other researchers outside their immediate project context is not especially important to most researchers. Much more important is developing ways to save time and/or resources while still achieving their project's scientific objectives and gaining positive community recognition for their work.

To be successful then, markup languages must be seen as making it easier to describe and convey raw data and results in complex contexts and support the reuse of those data and analyses. To accomplish this, markup languages must capture the dynamic nature of scientific properties, data collection, analysis, and methodologies and record the fullness of experiments and observations, as well as the associated modeling, simulation, and theory. This requires that markup languages be considered not in isolation, but rather in concert with formal, consensus-controlled vocabularies and ontologies. Markup languages need also to be flexible and robust enough to evolve over time in parallel to the dynamic nature of knowledge and predictive enough to support the long-term reuse of data. Markup languages have a potential to become adopted and used because data are rarely self-explanatory, and data exchange takes place over large space and time dimensions (e.g., we can't always predict at time of publication, the long-term cross-discipline use demands and different use models). To the extent that markup languages can better deal with such uncertainties and long-term objectives, they will be seen as beneficial and useful.

The plenary presentation concluded with a challenge for workshop participants to make markup languages address first and foremost the needs of scientists, seen as the primary users of scientific markup languages. And just as human languages evolve over time -- i.e., through the development of contractions, the reordering of syntax, the borrowing of words from other languages, the dropping and adding of grammar rules based on usage -- markup languages must be developed with an eye towards evolution. Nature is tricky; describing nature is even trickier.

Scientific Data Exchange Standards & Markup Languages by Domain

Following the introductory plenary, four brief presentations on the current state of scientific MLs as used in four specific scientific domains (chemistry, earth sciences, materials sciences and mathematics) served to highlight both the similarities and differences in how scientific MLs are being used in different disciplines, and to show the varying maturities of ML development by discipline. In the discipline-specific breakout discussions which followed, participants considered the following questions, which helped expand on the issues raised in the introductory presentations:

1. How can markup languages address, or be used to address, educational needs?
2. How does/will the domain-specific markup language(s) engage the user/domain community?
3. How does/will the domain-specific markup language(s) engage the international standards community?
4. Are there additional topics or issues specific to this domain-specific markup language(s)?

The presentations and breakout discussions repeatedly reinforced the idea that users of a given domain-specific ML approach the language differently according to their objectives. Scientists, as original generators of scientific output, have one set of needs centered on ease of authoring. Publishers of scientific content have a different set of needs and expectations of a markup language centered on management, distribution and display. And, end-users, both lay and specialist (and increasingly computer applications), have yet another set of expectations. For MLs to move forward in a discipline, adoption and development must occur among these user communities simultaneously. The next sections include a brief synopsis and discussion of the morning's plenary presentations and a summary of the major observations and outcomes of discussions by scientific domain.

Mathematics – MathML

Mathematics[7] benefits from having an established and pedigreed markup language (MathML) specific to the domain. Initial work on MathML predates even the formal release of XML as a W3C Recommendation and draws on early experiences from SGML (e.g., the ISO 12083 Mathematics DTD fragment) and HTML (e.g., the abortive effort during the development of HTML version 3 to augment HTML with a number of math specific elements, attributes, and constructs). MathML is an official recommendation of the W3C. Version 1 of MathML was released in 1998, and as of this writing the current release of MathML is version 2, second edition. Additionally, there is a relatively long tradition and established consensus within research and academic mathematics regarding standardization of notation, and in particular, a tradition encouraging the use of standards such as TeX for describing mathematical notation in digital form. There remain, however, a number of substantive issues with regard to MathML.

As one of the very first domain-specific implementations of XML, there were (necessarily) growing pains, and MathML is still seen as somewhat experimental by many potential users in the math community. Mathematics notation is extremely complex, with new notation required routinely to deal with cutting edge research in the field. MathML is therefore recognized as inherently incomplete. The authors of MathML have explicitly targeted it for the expression of mathematical content up through the early undergraduate level (first-order calculus). Its utility for research mathematics, even with its explicit built-in extension mechanisms (e.g., as exploited in the EU funded OpenMath project), is still uncertain. MathML is also intentionally bimodal, containing sets of elements to describe separately the presentation of mathematics and the semantics of mathematics. Generally, early implementers have focused on one or the other but not both parts of the ML, resulting in asymmetrical implementations that don't always interoperate as well as might be desired. Adoption has been somewhat slow, in part because of the entrenchment of TeX within the research mathematics community. Additionally, although mathematics is recognized as key to many scientific disciplines, and there have been some attempts to incorporate or accommodate MathML markup rules within other domain-specific markup languages, there are examples of domain-specific markup languages (outside of pure mathematics) that include their own markup semantics for basic mathematics needed within the domain of interest, rather than borrowing from MathML as needed.

On the positive side, there have been major inroads in getting the middle layer of the scholarly communication architecture to embrace MathML. While few if any research

[7] Robert Miner, Design Science, gave the introductory presentation regarding the state of markup language development in the discipline of mathematics.

mathematicians author directly in MathML (a surprising number do author directly in TeX or the related LaTeX language), publishers (e.g., the American Institute of Physics and the American Physical Society) and vendors of computer algebra engines and related tools (e.g., Wolfram Research, Design Science, Maple) are incorporating MathML into their workflows and products. This is being done to provide a high level of interoperability between systems and potentially to provide (in the long term) an enhanced user experience for the consumers of mathematical content. In his introductory remarks about MathML, Robert Miner identified four likely long-term benefits of broader adoption and use of MathML:

- facilitating cross-media publication, dissemination, and consumption of mathematics in an increasingly XML environment;
- providing another way that developers of authoring systems, digital libraries, and related applications can leverage their XML investments;
- meeting the legislative demand for better accessibility to content and educational resources containing mathematics; and,
- enabling more robust and general-purpose interactive mathematics tools.

The mathematics breakout discussion included a diversity of MathML experts and current and would-be users and consumers of MathML. This diversity of backgrounds and perspectives made for an energetic and wide-ranging discussion.

In discussing the potential benefits that MathML might bring to bear on educational services and models of learning, there were multiple points of consensus as well as several open issues and uncertainties identified. There was a consensus that MathML is increasingly being seen as an attractive *lingua franca* for expressing and sharing mathematical content in an educational context, at least among tool builders and digital library developers. Early experimentation and the successful integration of MathML in math education software like *Mathematica* and *Maple* and in tools for the Web made available from Design Science and Integre Technical Publishing give concrete evidence that the dissemination, interchange, and exploitation of mathematical content at a level appropriate for many educational purposes can be accomplished effectively using MathML. Interactive implementations designed for use by end-users in an educational context remain relatively immature and "hand-crafted" at this point in time, but show great promise. There is widespread demand by faculty at the middle school, high school, and undergraduate levels for better ways to put online for student use mathematically focused educational content, and correspondingly great interest in the potential of MathML to satisfy those needs. Several scholarly publishers as well are on record as seeing the potential of MathML to better support the dissemination of scholarly content containing significant mathematics, and a subset of scholarly publishers in science and technology domain are in fact facilitating the use of MathML through the creation and release into the public domain of specialized mathematical fonts and glyphs[8].

[8] The so-called STIX Fonts project, http://www.stixfonts.org/

That said there remain several open issues as well regarding the potential of MathML to help meet educational needs for a better way to express mathematics in online documents and learning resources. The utility of MathML to enhance searching and improve accessibility of online mathematical content has not yet been proven. Searching of mathematically laden content by the mathematics it contains is a complex issue. It's not altogether clear whether the level of description implicit in content (semantic) and/or presentational MathML is sufficient to support robust searching on the mathematics contained in a resource. It's also not yet certain that readers and other accessibility tools will be able to exploit MathML effectively to make the mathematics embedded in a resource more accessible, though that seems a safer bet. While MathML is being adopted (at least experimentally) behind the scenes -- e.g., as an exchange format for interoperation between applications like *Mathematica* and *Maple* and in the editorial workflow of scholarly journals, it has not been widely adopted by the authors of educational and scholarly mathematical content. Research mathematicians continue to rely heavily on TeX, which though exclusively presentation oriented (really a specialized language for the typesetting of mathematics) is firmly entrenched. Educators continue to rely on cruder technologies (e.g., embedding mathematics as static images within HTML or presentation only markup within PDF documents) or exploit proprietary solutions such as *Mathematica* workbooks. There remains a bit of a "chicken and egg" problem in that authors are hesitant to adopt a new technology until it has proven its value, and it remains difficult to prove the value of MathML without a sufficient body of MathML content.

Discussion of this issue led naturally into an extended discussion as to how MathML is now or might in the future engage the mathematics community. It is clear that MathML at this point in time is more appealing to organizations or institutions than it is to individual practitioners. As a non-proprietary, expressive, comparatively low-loss way to represent mathematics, MathML has clear attractions for long-term archiving and interchange of mathematics on a large scale. Hence its attractiveness to publishers and middleware tool developers. Several participants in the breakout session suggested that MathML may continue to develop as a largely or even exclusively back-end technology, used behind the scenes as a way to store and exchange mathematical content, but not necessarily as a format with direct impact on the author's or the end-user consumer's experience interacting with mathematical content. That would still make MathML useful, but the consensus was that MathML's greatest potential both economically and in terms of new functionality will not be realized until it is used more widely by content creators and ultimate consumers. This will require even more aggressive development of necessary authoring and presentation tools (including interactive presentation tools) and the inclusion of MathML within markup schemes developed by other science and technology communities that require the ability to express rich mathematics in documents and learning resources. This, in the collective opinion of those participating in the Mathematics breakout discussion, suggested avenues of common interests with other markup language communities represented at the workshop and led to the identification of several key issues of importance to the further development and future evolution of MathML:

- the need for more ubiquitous, more transparent (to the user) support for MathML in the Web environment;
- the need for better support within XML and Web-based applications for "compound documents" (i.e., as defined by the W3C, documents that combine multiple formats, such as XHTML, SVG, SMIL and XForms);
- better assurance that MathML will be maintained as a standard going forward;
- more sophisticated tools, especially on the authoring side, that can facilitate inclusion/embedding of MathML within online resources (e.g., within Web pages);
- continued development of better, more robust transformation tools (e.g., between TeX and MathML); and
- viable business models to better support and encourage ongoing development of MathML.

Earth Sciences – ArcXML, ESML, GML, NcML

The earth sciences, as a collection of domain sciences that include such examples as atmospheric science, oceanography, geology, have a particular set of needs to which the development of earth science related markup languages must relate. Needs include accommodating massive, and increasingly large amounts of data from multiple sources such as in-situ instruments, satellite remote sensing, observation campaigns (e.g. sonar sweeps of the ocean floor). These data are archived in a variety of ways from files managed by individual researchers to vast data archives managed by national authorities (e.g. national data centers). Collectively, the data represent terabytes of storage. Such data is collected as a one-time measurement and must therefore be preserved for later use.

In addition to the observational data, more data is being made available from models and simulations, often run on high performance computing systems, and capable of producing gigabytes of data per model run. This is especially so in climate modeling and in geodynamics modeling of the earth. In these cases, the models are often initiated from, and calibrated against, observational data. Here, the complexity of the models and simulations are important to record and preserve.

In this respect, the impact of data in the earth sciences in relation to markup languages relates to the opening session speakers remarks on developing markup languages to deal with observation data or experiments. In the case of Earth science, we can see a need for languages to deal with observation data (and thus the additional information related to the data (e.g. instrument reference and calibration, data provenance when data has been cleaned up), and to also deal with the model, or simulation applications as a form of experiment-based data.

Other common aspects of the Earth sciences beyond the collection and preservation of large amounts of data are the need for data to be geo-referenced and time-referenced. In some cases the requirement may be more specific to one, for example, geo-referencing of a location such as a city, or physical occurrence such as a rock type. In other cases, the need is for a complex 4D data model such as in weather and climate data. A result of the differences in needs is the development of a number of earth science related markup languages (e.g. the Geography Markup Language (GML) developed by the Open GIS consortium to enable descriptions of geometric objects and location, and the Earth Science Markup Language (ESML) developed by the University of Alabama at Huntsville as a dataset description language). A number of other format-specific languages (i.e. related to particular data archiving formats or analysis systems) have been developed (e.g. the NetCDF Markup Language (NcML) from the University Corporation for Atmospheric Research, or ArcGIS Markup Language (ArcXML) from ESRI, the makers of the ArcGIS software package).

Crosswalks and Mediation

In the cases of the languages mentioned so far, they all provide a syntax for the classification of information and data. Again, the importance of information related to data was highlighted in the opening session plenary with the observation that data are rarely self-explanatory. Self-documenting data standards such as NetCDF have emerged as leading formats in Earth Sciences (and beyond), and so the development of a NetCDF ML was a natural step.

The opening plenary also noted the need to realize that scientific language evolves over time, and that across domains, the language of science can be different. The need to work across domains is becoming increasingly important within the earth sciences along with the need for data integration. This requires not only understanding of the data (how it's collected, managed, units etc.), but also how this is used in the science of a given domain. As the data explanations and the language of the science (semantics) will be embedded in the syntactic structure of a given markup language, the need to have mechanisms to relate these will become important. Thus as the markup languages develop in specific communities and domains, we will need the means to map between them.

A first step at mapping is the development of crosswalks, a practice already well understood in the metadata world. In developing a crosswalk, the various elements in a language are mapped, in an agreed way, to elements in other languages. However, this can present problems when there is orthogonality between sections of languages, (i.e. in those areas that one language addresses, but another does not). The degree of orthogonality may represent those areas where a language may want to adopt from another if it makes sense. This approach is often cited as a means to lead to a common, canonical form in concept, but this may not work in practice due to the large degree of formal agreement on merging the semantics used in a language, and the ability of the stakeholders of one language to understand how others use, or extend their language. The markup language structure, and the mappings between language structures, could be made available through registries.

Another approach could be to allow many markup languages to flourish and to develop higher-level, or knowledge-level, mediation mechanisms between them. The idea of ontologies and concept spaces has been around for a while, but they are beginning to attract more interest in the area of possible mediation mechanisms. An example of an ontology in the Earth sciences is SWEET (Semantic Web for Earth and Environmental Terminology, developed at NASA/JPL) which is designed to enable the scalable classification of earth science concepts. SWEET is available as an OWL structure that can be used in tools being developed in the semantic web community such as inference engines. OWL (Ontology Web Language) is being developed by the W3C and is one of the technologies associated with the semantic web. The elements of the base markup languages can be registered with the ontology (e.g. SWEET semantic tags can be added to ESML descriptions) and thus relationships between languages can be inferred.

Visualization and Views

XML has provided a mechanism to develop machine-readable information objects where the syntax of the object is agreed upon to carry semantics that may not necessarily be agreed upon. The development of a markup language that builds on XML is that there can be structure agreement to the syntax as well as semantic agreement in how elements in the syntax are filled in. We have noted above that there are approaches that may allow us to go from one language to another, from simple crosswalks through using more complex knowledge models. However, a single language as expressed through its schema can be complex with resulting documents not being really human readable. There is a need for tools to not only build the documents, but to also view and use them. For example, a GML representation of the features of a city may embed a lot of properties about the features, but a user of the information may want to see a rendering of only a sub-set of features and properties.

The breadth of audience that may have a need to use a particular document expressed in a markup language can be broad; across scientific domains, from different educational levels of K-12 through undergraduate, graduate and informal. Each audience will need to access the language through a view that accommodates their needs. This also means dealing with presentation language issues. The semantics used in a markup language may be understandable at the undergraduate/research level, but would not be at the K-12 level where a different, perhaps simpler, set of semantics are needed.

Chemistry – CML

Chemistry Markup Language (CML) began in 1994 and, like the other markup languages represented at the Workshop, is based on XML from the World Wide Web Consortium. Built upon STMML (scientific, technical and medical markup language) used in publishing, CML is comprised of 5 parts: CMLCore (micro molecules, atoms, bonds) CMLSReact (reactions), CMLComp (computational chemistry), CMLSpectra (spectra) and CMLCryst (crystals). The early adopters of CML have been from some government agencies, such as the National Cancer Institute and the National Institute of Health as well as the European Patent Office. Some societies and publishers, like the Royal Chemistry Society and Nature Publishing Group, are committed to adoption of XML and CML.

The vision for the Chemical Semantic Web is an infrastructure where a robot can find phase diagrams for lipid mixtures or add molecular data to a researcher's monthly report following specified guidelines. Examples of more advanced applications would include reading a published paper in the *Journal of Medicinal Chemistry* and computing the geometries and energies for all new molecules, calculating binding to HIV protease, ordering the chemicals required for synthesis, checking safety, calculating and testing the chemicals.

Barriers to reaching a Chemical Semantic Web include intellectual property and economic (e.g., affecting publishers' and database providers' traditional business models) issues as well as resistance to change. A common theme with the other markup languages represented at the workshop was that components of markup languages are being built haphazardly by projects. The preference by project members is that some standards body or publisher coordinates this work. By targeting industries that are poised for innovation and that need to deal with large amounts of data, such as the pharmaceutical industry, CML and other markup languages can narrow the difference between advancing markup languages for profit as well as for research and discovery.

Currently most information associated with chemical work is destroyed when papers are published as PDF and Word documents. Word and PDF documents need to be transformed into XML and CML; however, publishers and secondary database providers are resistant to change traditional business models. Going forward, emphasis should be placed on creating compound documents that merge text and data, such as papers with editable chemical equations that animate chemical reactions. The development of authoring tools, browser enhancements, and generic physical science ontologies would enable early adopters, key industries, publishers, and software developers to work together in a coordinated effort to advance markup languages to meet real world needs.

Materials Science – MatML

Much of science and technology owe their progress to the careful collection, logging and interpretation of data. And as information technology becomes more efficient, so do the methods scientists use for sorting and accessing data. In order to improve the utility of electronic materials property data, the Materials Science and Engineering Laboratory of the National Institute of Standards and Technology with Ed Begley as the Project Leader, initiated and coordinated the development of Materials Property Data Markup Language (MatML) to standardize the way such information is posted and exchanged over the World Wide Web. The goal of MatML is to create a standard markup language for web-based materials property data collections in order to specify the hundreds of materials properties materials scientists and engineers need to know and access. By developing a markup language that describes the data source, the material, and its properties, the MatML effort aims to allow users in the industrial, research, and education communities to easily use and exchange electronic materials property data from multiple sources in models, simulations, or distributed databases.

MatML is a relatively new markup language designed for the exchange of property data values. With initial development begun at NIST in 1999, today MatML has broad representation and participation of private industry, government laboratories, universities, standards organizations, and professional societies from the international materials community. Motivated by the lack of a common materials data exchange format , as well as the tremendous opportunities for data exchange and dissemination supported by the World Wide Web, the MatML initiative grew out of two approaches: (1) general-purpose markup languages, e.g., Standard Generalized Markup Language,
(SGML), HTML, XML and,
(2) materials data standards, e.g., those promulgated over the past two decades by the American Society for Testing and Materials (ASTM), the International Standards Organization (ISO), and other standards organizations.

The MatML efforts are led by a Coordination Committee and is comprised two working groups: 1.) Schema Development Working Group and 2.) OASIS Standardization Working Group. The Schema Development Working Group has produced the MatML schema 3.1 (May 2004) aimed for use by those involved in the development, reporting, interchange, and application of materials information, including:

- Testing laboratories
- Database and software developers
- Information publishers
- Researchers and educators in materials selection from preliminary to final stages
- Designers using materials information (especially but not limited to FEM application)
- Materials quality assurance assessment

In today's global commercial materials community, a single company spread across the world needs to integrate a greater variety of specific data on demand to meet the requirements of efficient and cost effective product development and manufacturing. Additionally, the international materials research community, with collaborations across the world, is creating large scale, complex data more rapidly through sophisticated modeling, simulation, and experimental techniques such as combinatorial and multiscale analysis. The materials educational community seeks to prepare students for 21st century jobs in the nanotechnology sector by designing new curricula at the undergraduate and graduate level to educate and train the next generation of scientists and engineers, introducing them to the technical capabilities and skills needed in this highly interdisciplinary field.

Themes across Domains & Markup Languages

During the final wrap-up session of the workshop, representatives from each topical group (Education and Domain Experts; Markup Languages (in general); Publishers / Professional Societies; and Database / tool developers & data users) reported on common cross-domain issues that were discussed in the afternoon sessions. In the process of preparing this report, these issues were reviewed, along with key comments from morning presentations, and five themes emerged. These themes encapsulate the current state of activity and thought, not just on ML development and use by groups working independently, but also on the effects of ML's broader dissemination and use in the context of various sectors (e.g., education, publishing, government). These themes, despite the high-level tone of the titles, represent a significant step in the development of MLs. By reaching consensus on current challenges and opportunities, the various domains and MLs represented at the workshop can proceed on ML development with similar assumptions and priorities.

Theme A: Vision

Encapsulating information in XML underpins the interoperability concepts in the current web services environment where information, or data, encoded in XML can be easily exchanged between systems. As highlighted by the domain-specific breakout groups at the workshop, the development of markup languages that build on the XML framework (as standardized by the W3C) has generated a lot of momentum in the sciences over the past few years. Motivating the development of markup languages that are built on XML is the belief that by providing a means to exchange information, or data, in a structured form that colleagues across scientific domains can read, understand and use, scientific research and discovery can be moved forward. Through common interoperability mechanisms, NSDL supports the exchange of information between the sciences and provides a framework for markup languages to be extended even further as they are tested and applied in science education settings.

Theme B: Demonstrating the value of markup languages

Workshop participants returned on multiple occasions and in several contexts to the discussion of issues relating to the value (current and potential) of scientific markup languages. While there was a clear consensus (albeit largely intuitive and qualitative) that markup languages can be of significant benefit in scientific research and science education, it also was clear that likely benefits are spread across many different classes of markup language users, and that the case establishing that benefits outweigh start-up and ongoing implementation costs needs to be made better to users and to potential funding sources in order to stimulate broader adoption of scientific markup languages.

Participants identified issues common across domains that explain, in part, why scientific markup languages have not been more widely adopted. Markup languages have been established as a good way to link between information objects. However, despite the potential to benefit several science and research applications, their value in those contexts remains unproven. Markup languages' broadest implementation to date occurs in processes that are virtually invisible to most users. Specifically:

- The value of markup languages as a good, economical way to facilitate linking between information objects has been fairly well established. Similarly the value of markup languages as a good, economical way to transfer collections of document instances from machine to machine (interoperability) and to machine process (generically at least) collections of content is reasonably well established.
- Many if not most significant implementations of scientific markup languages have been undertaken in the context of backroom and "middlemen" applications in the scholarly information creation and dissemination cycle. One example involves several scholarly publishers (e.g., the American Institute of Physics, the American Physical Society, and Elsevier) who make extensive use of markup languages in in-house editorial and publishing workflows. Preliminary indications from these experiments by scholarly publishers are that markup languages, appropriately implemented, can greatly facilitate editorial processing, long-term storage, and reuse of formal published content such as journal articles, and can do so in a cost-effective way. To date though, use of markup languages in scholarly publishing has largely been a backroom phenomenon. Authors still submit in other formats, and content is still generally delivered by publishers in PDF, HTML, or other less structured (as compared to XML-based implementations) formats.
- Markup languages seem to have potential benefits in several other respects, but these have not yet been demonstrated systematically and/or in broad enough contexts to be considered proven in arena of science research and science education. There was a consensus that markup languages have potential benefits as an approach that would facilitate (and possibly enable): better, more precise searching of full-content; automatic extraction of object metadata (and other types of useful distillation of content); better support for compound document formats[9] and dynamic formatting generally; and, better facilitate scholarly dissemination tasks such as peer review, versioning, and archiving.

These potential benefits are seen as common across most science markup languages. Additional research is required to better quantify potential benefits (and costs) of markup languages and establish more persuasive cost-benefit models.

Theme C: Creating & disseminating the pre-requisite tools

Participants from all domains agreed that better tools, both technically and in the form of broader, more robust ontologies, would facilitate and speed the adoption of scientific markup languages. Specific examples included:

- Better, standardized, and preferably, open source tools are needed to transform and translate between various scientific markup language dialects. For instance it would

[9] In the W3C sense of the phrase; see: http://www.w3.org/2004/CDF/

be desirable to be able to extract and reformat in MathML the mathematics contained in a CML or MatML document.

- Validation tools that can more effectively vet the accuracy and internal consistency of information objects and the viability and correctness of links to external references.
- Better rendering engines that accurately and consistently render marked up content and are capable of better exploiting compound document and dynamic formatting features.
- Better, more comprehensive ontologies (and associated cross-walks) adequate to support a greater level of autonomous processing of marked up content and more robust search and discovery and interchange of information across disciplines.
- Tools capable of extracting and utilizing in a more autonomous way the implicit semantics of XML-based marked up content.

Even as these needs were raised, examples of currently available local or discipline-specific tools that address some of these needs (at least in part) were also brought up. Clearly there is an extensive body of prior art in the arena of markup language tools, but just as clearly, few if any participants had a clear view of the current landscape of available tools and ontologies. This suggests two specific actions:

Theme D: Mediation of markup languages

The need for cross-markup language understanding (e.g. how do the structure and semantics of one language relate to those of a second language) was a common theme in the workshop, mentioned in the keynotes and various breakout discussions. While workshop participants did not use a specific term, the report editors decided that "mediation" best covers the concept of tools and services that provide a translation interface between representations in different markup languages, or that provide access to information in a single markup language to a wide variety of users.

The translation may be needed at a human level, or at a machine level where two systems may need to interoperate. The development of taxonomies and controlled vocabularies are one approach to the cross-language interoperability. Ontologies are another approach to being able to relate the semantic meaning of one language to that of another. A number of the markup languages discussed at the workshop have rich vocabularies in their structures, but the idea to link these to a broader knowledge representation, such as an ontology, are just beginning to be explored.

Another aspect of mediation relates to end user access to information as it is structured in a particular language. For example, the semantic structure of a language has been developed by early developers, often in the research area the language is directed toward. Thus, the end user must have a high level of domain understanding in order to use, or access, a document in the language. This has implications for how end users who are new to, or unfamiliar with, a field make use of the content instantiated in a markup language. For NSDL, with its mission to support education, providing access to markup languages for users who are not domain experts is critical. Thus, a second aspect of mediation in relation to end users is in the development of tools and systems that can provide access to documents, or data, held in a markup language such that non-experts in the field can understand and use the information. For markup languages to be used widely in education, it will be important to develop tools and services that can mediate the semantics of markup languages to pedagogical concepts related to developing domain understanding.

Theme E: Identifying challenges to maturation of markup languages

Poorly developed scientific markup languages run the risk of being obstacles to useful interoperability and impediments to innovation – the exact opposite of their intended function. A common issue in the growth of markup languages is the tension between reflecting the dynamic nature of science (and thus the ever-changing landscape of the scientific markup language) while supporting a certain amount of standardization, which cements a language but which also enables broad interoperability. Workshop participants' consensus was that this tension can be successfully addressed only if a broadly consultative and inclusive language development process is given an adequate amount of time and support. To find the right balance between fixity and flexibility requires an iterative cycle that is open to contributions from a diverse community of experts representing not only scientists, but also publishers, librarians, educators, students and other end-user consumers of scientific research. There is a cultural challenge to sustaining an attenuated consensus-building process in that many scientists are conditioned to be entrepreneurial and are not often required to involve professionals outside their immediate peer group and domain of expertise. However, several scientific markup languages have now reached a critical juncture in their development where broader input, more thorough testing, including software implementations, and development of consensus involving publishers, educators, and end-users is required to insure proper maturation.

NSDL 2004 Annual Meeting Panel

A panel discussion was held at the NSDL 2004 Annual Meeting as a follow up to the Workshop. Given the importance of discovery and reusability in the development of effective learning resources within the NSDL community, the panel provided the platform to discuss 1.) how markup languages support discovery and reusability and 2.) how markup languages can help enhance end-user interactions with science content to advance innovative science education. Discussion among the audience and other members of the panel about using markup languages in the NSDL highlighted the following points:

- FEDORA's generalized object-to-object relations can support applications using markup languages.
- NSDL is well suited to re-energize work on UnitsML because many NSDL projects would benefit from the effort.
- There is a need for education on ontologies and their benefits/uses in NSDL.
- Future workshops should expand the educational component.
- Use of MLs for rich interfaces can conflict with designers' wish for simple interfaces.

Participants also noted a market-related challenge to the maturation of scientific markup languages. Traditionally, commercial self-interest provides adequate incentives for publishers to convene the necessary experts to write, test, and promulgate a standard that insures interoperability. However, since the means to implement XML-based markup languages standards are non-proprietary and transparent by definition, the business models for initial software implementations based around scientific markup languages are marginally profitable. At the same time, the intellectual effort for routine tasks involving markup languages is often greater than similar tasks as currently performed. For example, while an editorial assistant with an English degree might be adequately trained for processing STM journal articles to be published solely in print, that same assistant likely would need additional training (implying a higher salary after training) to encode an article in a sophisticated scientific markup language for dissemination online. Some estimates suggest such a staff upgrade across the STM publishing community could increase editorial costs as much as 20%.

The lack of viable business models and the prospect of increased operating costs become significant dis-incentives for commercial software vendors and publishers to convene a broader community of stakeholders required to develop a consensus around scientific markup languages and core software tools. Workshop participants recognized a need for quantitative cost-benefit cases that demonstrate the ultimate value of scientific markup languages (see Theme B: Demonstrating the Value of Markup Languages). But they also expressed concern that in the meantime, the evolution of some scientific markup languages might slow to a hazardous degree for lack of aggressive software implementations and language development. In this case, the government, rather than the private sector, might best support scientific markup languages during their transition towards maturity.

Conclusion

While it was recognized that there are problems identifying connections between domains, the commonality across the science community that overlaps multiple disciplines also provides avenues for opportunities. Throughout the discussions at the workshop as well as later interactions by e-mail and at the NSDL Annual Meeting, workshop participants focused upon identifying common, overarching workshop themes with associated recommendations that were relevant across scientific education and research communities. The consensus of the workshop participants was to begin with these common themes and recommendations as a way to move forward. The workshop results suggest the following:

1. NSDL has a unique opportunity (and an obligation) to continue to support cross-domain work on scientific markup languages

NSDL should be supported as a meeting point for scientific domains to further research on scientific markup languages. Specific immediate actions include

- Continuing the listserv supporting the workshop
- Develop a registry of scientific markup languages
- Plan a follow-on workshop

2. NSF should support the next stage of scientific markup language standardization and implementation

NSF, and NSDL, should take immediate steps to identify and convene the communities of experts required to take the most promising scientific markup languages to the next level. NSF should also expand targeted support for applied research to develop and prove business models and templates leading to the development and sustenance of software creation and staff improvements critical to the broader adoption of scientific markup languages.

3. NSF / NDSL should support and encourage increased collaboration across and among domain-specific scientific markup language communities, e.g.:

3.a. Conduct an environmental scan of scientific markup language tools and ontologies

NSF, possibly in conjunction with other agencies with coinciding interests (e.g., NIST), should undertake an environmental scan and inventory of currently available scientific markup language tools and ontologies, with a longer range goal to engender a formal or informal registry of such resources.

3.b. Support research on mediation services and tools that operate between markup languages

NSF should support further research into the development of mediation services between markup languages that would incorporate more formal knowledge representation techniques such as ontologies.

3.c. Support research on generic services and tools that mediate between scientific markup languages and end users in education

NSF and NSDL should support research that investigates and develops tools and services that will allow a wide range of educational users to access the information and data described in a markup language framework through innovative use of learning structures such as concept maps and education-based knowledge organization systems.

3.d. Work with appropriate organizations to encourage to conclusion the development of UnitsML.

Markup for scientific units is a pervasive and long running problem. NSDL should work together with the interested stakeholders to advance a cross-language solution to units, which includes establishing a UnitsML working group under the NSDL Technology Standing Committee.

4. NSF / NSDL also should support and encourage targeted, applied research to:

4.a. Better assess and quantify the potential benefits of markup languages (singularly and generically)

NSF and the NSDL program should consider encouraging and supporting more systematic and quantitative research that assesses potential benefits (described above) of scientific markup languages and that develops better metrics for measuring the quality of markup language design (i.e., in terms of potential benefits to authors, intermediaries, and ultimate end-users).

4.b. Produce better tools and ontologies for use in concert with scientific markup languages

4.c. Identify additional target audience(s) which can benefit from scientific markup languages

NSF should consider funding one or more needs assessment research projects to better identify primary audience(s) that will benefit to the greatest degree from more extensive and complete implementations of one or more of the emerging scientific markup languages. Some questions to address include: What problems do scientific markup languages resolve, and for whom? What education levels will benefit? Do faculty need to bring some level of expertise, either domain-specific or XML knowledge, to reap the benefits of scientific markup languages?

Appendix A: Presentation & Demonstration URLs

All of the documents association the workshop can be found on the workshop website: http://scimarkuplang.comm.nsdl.org/. Below are URLs from the plenary presentations.

Plenary Presentations

- Workshop Reception Presentation, 14 June 04
  Making the Web Safe for Intelligent Agents
  Professor Tim Finin, University of Maryland Baltimore County
  http://ebiquity.umbc.edu/v2.1/resource/html/id/32/

- Workshop Opening Presentation, 15 June 04
  The Dynamics of Data Standards
  John Rumble, Information International Associates
  http://comm.nsdl.org/download.php/357/rumble-SciML_presentation.ppt

Presentations on Domain-specific Needs for Markup Languages

- *Math ML: An Overview*
  Bob Miner, Design Science
  http://comm.nsdl.org/download.php/362/miner.zip

- *Markup Languages - Earth Systems Science*
  Rob Raskin, NASA Jet Propulsion Laboratory
  http://comm.nsdl.org/download.php/358/raskin.ppt

- *Chemistry Markup Language*
  Peter Murray-Rust, Cambridge University
  http://comm.nsdl.org/download.php/364/peter_zip.zip

- *Using Mat ML in Research and Education*
  Adam Powell, MIT
  http://comm.nsdl.org/download.php/359/powell.ppt

Presentations prepared for panel of breakout discussions

- *Scientific Markup Languages: Implications for Publishers*
  Tim Ingoldsby, American Institute of Physics
  http://comm.nsdl.org/download.php/360/ingoldsby.ppt

- *Scientific Markup Languages Interoperability*
  Stefano Nativi, University of Florence
  http://comm.nsdl.org/download.php/361/nativi.ppt

Appendix B: Workshop Agenda

Monday, June 14, 2004

6:30 - 8:30 NSF / NSDL Workshop Opening Reception
Hilton Garden Inn, Arlington/Courthouse Plaza, Chambers 1 & 2

7:00 Welcome and introductions

7:10 Making the Web Safe for Intelligent Agents
Professor Tim Finin, Computer Science and Electrical Engineering
University of Maryland Baltimore County
7:40 Q&A

(heavy hors d'oeuvres and cash bar; attire is business casual)

Tuesday, June 15, 2004

8:30-4:30 NSF / NSDL Workshop on Scientific Markup Languages
NSF Building, Room 375

8:30 Continental Breakfast (Vegan/Atkins selections)

9:00 Session I - Plenary: Welcome & Presentations

9:00 Workshop Introduction
9:10 Welcome from NSF – Lee Zia, NSDL Program Director
9:20 An Historical Perspective on Markup Languages
John Rumble, Information International Associates

9:45 Overview of Domain-specific Needs for Markup Languages

9:45 Math - Bob Miner, Design Science
10:00 Earth Sciences - Rob Raskin, NASA Jet Propulsion Laboratory
10:15 Chemistry - Peter Murray-Rust, Cambridge University
10:30 Materials Sciences – Adam Powell, MIT

10:45 Wrap-up; Break

11:15 Session II - Breakout: Completing the Picture of Domain Needs

12:30 Lunch

1:15 Session III - Breakout: Complementing & Extending Our Work

3:00 Break

3:15 Session IV - Panel: Wrap up Discussion

4:30 Workshop Ends

5:00 Shuttle leaves for Hilton Garden Inn

Appendix C: Workshop Details

Goals for the Workshop on Scientific Markup Languages

- Begin discussion within and across domains and markup languages
- Produce a report describing workshop discussions and suggested next steps
- Identify the appropriate process and format for continuing and broadening discussions within and external to the NSDL community

Session I – Plenary (9:00 – 11:00)
The workshop will open with a plenary session that will provide participants with a framework for the rest of the day's discussions. Presentations will describe an historical perspective on the technical and social challenges associated with exchanging data across domains and will identify needs specific to the markup languages and domains represented.

Session II – Breakout (11:15 – 12:30)
Meeting in language-specific working groups, participants will respond to the issues raised in the morning plenary presentations to further identify and describe the current state of markup language development and use.

Questions to guide discussion:

- How can markup languages address, or be used to address, educational needs?
- How does/will the language engage the user/domain community?
- How does/will the language engage the international and standards communities?
- Are there additional topics or issues specific or special to a markup language?

Outcome: Groups will develop recommendations on the broader application and next steps for development within each language.

Session III – Breakout (1:15 – 3:00)
Participants will self-select into working groups around topics (described below) to discuss how to complement and extend work in markup languages and domains.

Outcome: Specific outcomes will depend on topics, but all groups are encouraged to identify challenges shared by all domain or markup language representatives prioritize them and develop short- and long-term action items. Similarly, groups should identify opportunities to complement or extend ongoing work.

> *Education and Domain Experts*: Discuss those issues relating to use of MLs to facilitate exploitation of resources in educational and domain context. Address questions such as: How will broader use of MLs within NSDL, or other digital libraries or sites, enable new educational or domain specific services or capabilities? Session will touch on research dealing with online learning styles and models, domain user needs.
>
> *Markup Languages (in general)*: Discuss those issues relating to markup in general which can inform design, use, cross-language mapping strategies, and documentation of specific Sci-Tech MLs. Address questions such as: How

do we improve quality and interoperability of Sci-Tech MLs by drawing on lessons learned from other language-specific and cross-language ML projects? Session will touch on issues of modeling, semantic interoperability, schema design, and documentation.

Publishers / Professional Societies: Discuss those issues relating to markup in a publishing context. Address questions such as: How can broader use of MLs facilitate development of better and more useful online publications? Session will include discussion of experience to date with publication-oriented DTDs and schemas and lessons learned from cross-publisher initiatives such as CrossRef / DOI.

Database / tool developers & data users: Discuss those issues relating to markup tools and use of markup languages in conjunction with databases. Address questions such as: How useful are emerging Sci-Tech MLs as data modeling approaches for applications like those being developed by NSDL? Session will touch on issues relating to migration of XML resources into and out of relational databases and other database structures and on the state of the art of current software tools used to manipulate XML data structures.

Session IV – Panel (3:15 – 4:30)
The workshop will conclude with a panel discussion:

- Panelists will briefly report results from Session III
- Panelists will discuss a process and possible next steps for implementing action items or continuing to extend work on domains / markup languages
- Panelists and audience will discuss the value of continuing and broadening discussions begun at the workshop with NSDL
- Panelists will discuss the role of markup languages in the emerging cyberinfrastructure.

Appendix D: Participant Statements

A .pdf of the statements is available on the workshop website.
http:// comm.nsdl.org/download.php/363-part-statements.PDF