Y/IT-332

B&W Y-12 technical services LLC

Y-12 NATIONAL SECURITY COMPLEX

Report of Official Foreign Travel to Montréal, Canada 1–7 August 2010

28 August 2010

James David Mason
SAIC

Babcock and Wilcox Technical Services Y-12, LLC
P. O. Box 2009
Oak Ridge, TN 37831-8169

National Nuclear Security Administration
PRIDE
Product Realization Integrated Digital Enterprise

MANAGED BY
B&W Y-12, LLC
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

UCN-13672 (1-08)

Report of Official Foreign Travel to Montréal, Canada 1–7 August 2010

28 August 2010

James David Mason
SAIC

Prepared by
Babcock and Wilcox Technical Services Y-12, LLC
Management & Operating Contractor
for the
Y-12 National Security Complex
under contract DE-AC05-00OR22800
with the
U.S. Department Of Energy
National Nuclear Security Administration

ABSTRACT

How can DOE, NNSA, and Y-12 best handle the integration of information from diverse sources, and what will best ensure that legacy data will survive changes in computing systems for the future? Although there is no simple answer, it is becoming increasingly clear throughout the information-management industry that a key component of both preservation and integration of information is the adoption of standardized data formats. The most notable standardized format is XML, to which almost all data is now migrating. XML is derived from SGML, as is HTML, the common language of the World Wide Web.

XML is becoming increasingly important as part of the Y-12 data infrastructure. Y-12 is implementing a new generation of XML-based publishing systems. Y-12 already has been supporting projects at DOE Headquarters, such as the Guidance Streamlining Initiative (GSI) that will result in the storage of classification guidance in XML. Y-12 collects some test data in XML as the result of Electronic Data Capture (EDC), and XML data is also used in Engineering Releases. I am participating in a series of projects sponsored by the PRIDE initiative that include the capture of dimensional certification and other similar records in XML, the creation of XML formats for Electronic Data Capture, and the creation of Quality Evaluation Reports in XML.

In support of DOE's use of SGML, XML, HTML, Topic Maps, and related standards, I served 1985–2007 as chairman of the international committee responsible for SGML and standards derived from it, ISO/IEC JTC1/SC34 (SC34) and its predecessor organizations; I continue to belong to the committee. During the August 2010 trip, I co-chaired the conference *Balisage 2010*.

Note: This report continues a series, the most recent of which, Y/ACT/FTR-224, reported on the conference *Balisage 2009* in Montréal, Canada.

INTRODUCTION

Over the course of the past two decades, SGML (Standard Generalized Markup Language, ISO 8879:1986) and its applications, including HTML (Hypertext Markup Language), and profiles, most notably XML (Extensible Markup Language), have come to dominate the interchange and use of structured data. SGML and many of the standards related to it and XML were developed and are maintained by ISO/IEC JTC1/SC34 (SC34), in which I have participated since 1981.

SGML- and XML-based publishing systems have been developed and deployed at numerous DOE and NNSA facilities for more than fifteen years. The most recent efforts at Y-12 are currently in progress with the Arbortext Epic system from PTC. The Arbortext system is also being installed at Headquarters and at other field sites, including Sandia and Lawrence Livermore, for managing classification guidance and other documentation.

XML is the basis of a series of projects sponsored by the PRIDE initiative that include the capture of dimensional certification and other similar records in XML, the creation of XML formats for Electronic Data Capture, and the creation of Quality Evaluation Reports in XML. The common element in all these efforts is the capture of certification and evaluation data in XML so that it can be used for more than the single purpose of creating paper records of the product life cycle. XML data formats offer the potential for utilizing the captured data directly in further analysis by both Y-12 and the design agencies.

CONFERENCE: *BALISAGE 2010*

Balisage is an international conference, now in its third year, and successor to the conference *Extreme Markup Languages* that had for the previous eight years focused on the cutting edge of development related to XML and other markup technologies. (*Balisage* is French for markup.) The conference has been sponsored by OASIS, a standards body; the World Wide Web Consortium (W3C); the Text Encoding Initiative (TEI); several corporate sponsors; and a number of XML users groups.

This year's conference opened with a one-day International Symposium on "XML for the Long Haul: Issues in the Long-term Preservation of XML." As XML applications expand in usage and are embedded in enterprise information technology architectures, they are processing increasing amounts of data. As a result, the viability and sustainability of XML data are becoming significant issues for system designers and implementers.

The regular conference featured three and a half days of papers that covered a wide range of XML applications, ranging from industrial documentation to XML processing to semantic analysis. As a member of the conference staff, I spent much of the time chairing sessions, so I did not hear the full range of papers being presented. I have, however, read the submitted review drafts of all the papers; the list of literature collected has links to the complete conference proceedings.

INTERNATIONAL SYMPOSIUM ON XML FOR THE LONG HAUL: ISSUES IN THE LONG-TERM PRESERVATION OF XML

As XML becomes increasingly dominant for data transfer and storage, ever-growing bodies of XML data are flowing across networks and accumulating in repositories. Coping with vast quantities of data has inspired approaches to ensuring that the data can survive and be useful over long periods of time. In some cases, where XML is being used to preserve data of cultural significance, concerns involve periods in the centuries.

Long-term preservation of XML data has some issues specific to XML, such as the roles of schemas and stylesheets and the entity structure of XML documents. However, many of the issues that affect XML also affect other forms of electronic archives. Such issues may involve details of data representation, such as character encoding and UNICODE versions, or they may be at a very high level, such as the need to maintain and refresh recording media. Almost all electronic records issues also involve metadata, and regardless of the nature of the records in an archive, the best practices for metadata generally involve XML.

The Symposium paper most closely related to the needs of Y-12, the NSE, and DOE/NNSA was "Metadata for Long Term Preservation of Product Data" by Joshua Lubell of NIST. Much of Lubell's paper was devoted to manufacturing data, for which NIST promotes the STEP standard for 3-D design. However, Lubell stressed that the success of maintaining product data depends on both the overall design of an archive and the metadata used in it. For archive design, he pointed to ISO 14721, Open Archival Information System (OAIS). For the XML metadata, he mentioned the Metadata Encoding and Transmission Standard (METS) schema and profile and the Preservation Metadata: Implementation Strategies (PREMIS) data dictionary and schema. He also discussed DoD's Technical Data Packages (TDP, MID-STD-31000) for collections of product data. For long-term retention and indexing, he recommended combinations of Dublin Core (DC) and RDF/OWL metadata.

Two papers presented views of maintaining technical information from the biological sciences. PubMed Central (PMC), part of the National Library of Medicine in the U.S. National Institutes of Health, is a rapidly growing (by 10,000–15,000 items a month) collection of several million articles related to medicine and public health. PubMed Central emphasizes bringing current research to a wide community. A complimentary service is provided by Portico, a digital preservation service for electronic journals, books, and other content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Much of Portico's archive is "dark"; that is to say, Portico is building collections that may not become visible unless their current manifestations, perhaps as offered by a publisher or journal, cease to be available.

Jeff Beck's "Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles" explained a strategy for developing a long-term archive. PMC receives articles from journal publishers in a variety of XML article formats (some are actually still SGML) and then converts them to its

own XML format that is used for online distribution. Beck again discussed the quality-assurance process necessary to maintain this conversion and assure that it adequately reflects the source documents (he has spoken in detail about this subject at previous conferences). In something over a decade of operation, PMC has already had to address long-term issues and has developed guidelines and documentation for what they do. (For example, their XML application has evolved over time, and several years ago they had to address the issues raised by a new version that could no longer maintain full backwards compatibility with earlier archives.) PMC is also looking at the possibility of the disappearance of some print sources and the issues raised by documents that are continuously revised, rather than being issued at a single point in time.

John Meyer presented "Portico: A Case Study in the Use of XML for the Long-Term Preservation of Digital Artifacts," which involves a strategy for what he calls "enduring usability, authenticity, discoverability, and accessibility of content over the very long-term." Portico's archive is even larger than PMC's, involving over 15 million items that comprise over 175 million files. While Portico is by design coordinated with PMC, their role as what might be thought of as an "archive of last resort" to guard against the failure of other resources causes them to build a complex infrastructure of metadata for preservation. They estimate that they accumulate about 100 gigabytes of metadata for every terabyte of content (and they currently have about 15 terabytes of content). Unlike PMC, which converts source data to their internal XML for distribution (while keeping copies of the sources), Portico intends to serve up the originators' XML (as well as offering versions converted to PMC's format). This means that Portico must do quality assurance on the source files and ensure that the source is complete (schemas, entities, graphics, and other attachments) and accounted for in their metadata. Meyer cited the paper that Jerome McDonough presented at *Balisage 2008*, "Structural Metadata and the Social Limitation of Interoperability: A Sociotechnical View of XML and Digital Library Standards Development," as a significant source of guidelines for dealing with heterogeneous sources of XML and metadata.

One of the ongoing themes of the conference series is the *meaning* of markup. Without getting into the more philosophical discussions in the main conference, the Symposium papers kept turning to this subject. As Meyer suggested, particularly by reference to McDonough's earlier paper, it is not always sufficient to have a schema when interpreting XML markup. The schema may allow validation of structure, but it does not necessarily enable understanding of intent. Two Symposium papers took on the subject of meaning: linguists are particularly concerned for the meaning of utterances, and that includes markup both for its own sake and for its ability to encode the structures of linguistic resources.

"Sustainability of Linguistic Resources Revisited," presented by Oliver Schonefeld and Andreas Witt, from the Institute for the German Language (IDS), Mannheim, looked at a number of approaches to markup. Like many working in literary and linguistic research, the authors made frequent reference to the XML supported by the Text Encoding Initiative (TEI). They found that they must supplement the TEI encoding with additional metadata, and, like PMC, must develop a scheme for normalizing and regularizing diverse markup. Schonefeld and Witt examined other issues as well. They had a particularly interesting section on the interpretation of OOXML, the encoding behind Microsoft formats like the ".docx" used by recent versions of Office. OOXML is a particularly complex and unconventional application of XML, and they observed that "just having the data encoded in XML does not automatically

make data sustainable." The authors also considered topics including "availability and findability" of both data and supporting resources once the data is encoded in XML. Like the other presenters at the Symposium, they also discuss the importance of "selection and qualification for long term archiving."

A second linguistics-based paper, "A formal approach to XML semantics: implications for archive standards," by Andrew and Quinn Dombrowski, of the University of Chicago, approached very formal structures (though with amusing examples) and offered suggestions for designing schemas for archives and metadata.

Most of the themes of the other papers came together in a concluding paper by Liam Quin, who leads all XML activities at the Wold Wide Web Consortium (W3C): "Beyond Eighteen Wheels: Considerations in Archiving Documents Represented Using the Extensible Markup Language." Quin observed that although the "very long time" which an archive may be called upon to serve may begin in the present, "what is significant is that the people who created the document are not the people who decode it, and that the context of that decoding is not necessarily the same as the social, technological or political context in which the document was encoded." Quinn's paper provided an excellent summary of themes from the others (though it was written before the full texts of the others were available). One particular issue that he raised is the need in archival documents to resolve shortcuts that may be useful in the authoring process but which might cause ambiguity in the later use of the documents. For example, Y-12 may use XML shortcuts to handle repeated references to the subject of a specification or to pull in boilerplate text; these need to be replaced with the full text upon completion of the document.

The afternoon of the Symposium was devoted mainly to a user forum.

BALISAGE 2010

Balisage 2010
The Markup Conference

This year's Balisage conference was the largest since the inauguration of the series. The conference continued a number of themes that have run through the series: What does markup mean? How can multiple layers of annotation be applied to a single document? How can systems best be made usable and useful for end users?

Because this conference coverd the whole range of markup languages and applications, a number of the papers were not directly relevant to Y-12. From those that were relevant I have selected several to feature in this report.

General Markup Practice and Standardization

The conference opens, by tradition, with a keynote by the overall conference chair, B. Tommie Usdin (Mulberry Technologies). This year she chose the topic "The high cost of risk aversion." Her theme was that taking what may in the short term appear to be the least risky approach to application development and implementation may in the long run expose both developers and users to the greatest risk.

Eric Freese (Aptara) started the technical papers with a look at "Multi-channel eBook production as a function of diverse target device capabilities." As the market for eBooks grows (Amazon has reported that they sell more eBooks than hardbacks) and the devices become more diverse—dedicated readers (Kindle, nook), smartphones (iPhone, Android), tablets (iPad), in addition to conventional computers—so does the need for flexibility in generating the eBook files. While DOE is not yet in the eBook business (OSTI might find some demand, however), the problems encountered in eBook creation are relevant to those creating XML documents for the long term.

Karen Wickett (University of Illinois) examined some issues related to the long-term use of XML documents in her paper "Discourse situations and markup interoperability." Like the Symposium speakers, Wickett emphasized the importance of context, for both documents and metadata, in the interpretation of markup. She looked particularly at Dublin Core metadata and its role in the OAIS model for archives. While traditional XML application design, going back to XML's roots in SGML, has tended to create element names from natural languages in the hope of capturing semantics for human users of the markup, this is not guaranteed to be successful, and even a widely used application like Dublin Core can be applied with unfortunate results.

XML applications can be built in many ways. In his paper "I say XSLT, you say XQuery: let's call the whole thing off," David Birnbaum (University of Pittsburgh) looked at two common languages for extracting information from documents. An application he had written some years ago in XSLT to graph the interrelationships among members of a corpus of mediaeval Russian manuscripts now looks easier to write in XQuery. Choice of application language may be affected by the nature of the task, but it is often a matter of the developer's style and taste. (At this point, most of Y-12's applications use XSLT, but as we develop archives we may need to increase our use of XQuery.)

Michael Kay (Saxonica), last year discussed his preliminary analysis of how much of XSLT 2.0 will be streamable. This year he discussed how he has implemented streaming in his Saxon XSLT processor (Y-12 uses Saxon in its applications). While Y-12's corpus of XML data is still small, and the files are not yet large, streaming might be a useful technique as future data-capture systems accumulate increasing amounts of data.

As libraries of XML documents increase in size, it is likely that multiple documents will share dependencies on certain components, ranging from schemas and stylesheets to boilerplate text, that must be managed, including version control. Y-12 is already on the verge of needing such support. Florent Georges (H2O Consulting) proposed a packaging system based on implementations of the EXPath extension to the W3C XPath standard: "The EXPath Packaging System A framework to package libraries and applications for core XML technologies." His implementation is based on some tools we already use, such as Saxon.

Two other papers, which I was not able to hear, might be of use to Java programmers: "gXML, a new approach to cultivating XML trees in Java, " by Amelia A. Lewis and Eric E. Johnson, and "Java integration of XQuery — an information unit oriented approach," by Hans-Jürgen Rennau.

Markup Theory

Balisage and its predecessor conferences have a long tradition of examining the theoretical underpinnings of markup languages and document markup. This year's conference continued the practice of looking at the nature of markup and even of documents themselves.

To begin with the most radical theoretical proposition: Last year Allen Renear and Karen Wickett (University of Illinois) proposed that "Documents Cannot Be Edited." This year they revisited the issue and concluded that "There are No Documents." By this assertion they are not attempting to be absurd but rather to examine the foundations of what it means to create and store documents, of what the contents of archives are. From this they conclude, in effect, that a document is a social construct, and that position is not unlike the position taken in the Symposium as well as in Wickett's individual paper, that a document depends on its context.

On a slightly less epistemological level, Claus Huitfeldt (University of Bergen), Yves Marcoux (Université de Montréal), and Michael Sperberg-McQueen (Black Mesa Technologies) considered an "Extension of the type/token distinction to document structure." The notion of a "type" is fundamental in XML, which has its roots in "generic" or "generalized" markup, particularly in the ISO standard for the Standard Generalized Markup Language. One of the fundamental concepts in SGML, carried over to XML (and used in most Y-12 XML applications) is a "Document Type Definition" (DTD). The usual sense in creation of XML documents is that they are *instances* of the grammar declared in a DTD. Problems with instantiation lie at the root of the conundrum proposed by Renear and Wickett. Huitfeld, Marcoux, and Sperberg-McQueen propose instead of instantiation that concrete XML objects, from documents down through elements to atomic characters be considered tokens, and that tokens have a property that can be called "type." Such a view of XML tokens is useful, for example, in understanding the nature of the objects in an archive.

The DTD, particularly as introduced originally in SGML, was powerful, though sometimes unwieldy for defining grammars for publishing applications. It was less useful for database applications, and it was less than satisfying for some computer science formalists. As a result, the DTD was modified in XML, and other schema languages were developed by ISO and the W3C to address some of the perceived shortcomings of DTDs. Maik Stührenberg (Bielefeld University) presented "Refining the Taxonomy of XML Schema Languages. A new Approach for Categorizing XML Schema Languages in Terms of Processing Complexity," which set out to extend the work of Makoto Murata, himself a developer of schema languages.

One of the themes that returns regularly in this conference series is multiple layers of structure in single documents. While the basic model of an XML document is a single-rooted tree, sometimes it is important to add layers, as for annotations. At Y-12 we have an interest in such concepts for projects like EBH, where an inspector might want to add free-form annotations to a form during a part certification. Pierre-Édouard Portier (Université de Lyon) presented a model for just such annotations in "Multi-structured documents and the emergence of annotations vocabularies." While the specific subject of the paper was annotation of scanned images of a mathematician's notebooks, perhaps the graphical user interface to the annotation system could be adapted to other applications.

Another approach to multilayered documents was presented by Hugh Cayless (New York University) in "On Implementing string-range() for TEI." Some XML applications compensate for the single-rooted nature of XML trees by using stand-off markup, which is to say they supplement the embedded markup of normal XML element tags with pointers into the file. TEI includes a "string-range" function that nonetheless remains unimplemented. Cayless suggested methods for implementing it. Such techniques might be of interest to Y-12's archives because they could support noninvasive means of annotating files that cannot be modified.

Markup Practice

This year's conference had fewer application case studies than previous years. Nonetheless there were some interesting papers.

Stefanie Haupt and Maik Stührenberg (Bielefeld University) examined a problem familiar at Y-12, constructing a rich XML application from source materials that are poorly or inconsistently marked up and lack direct semantic information about their contents. "Automatic upconversion using XSLT 2.0 and XProc: A real world example" took as its subject matter a collection of online reviews of video games. HTML is not semantically rich, and when it is used erratically it is a difficult starting point for data mining. (How, for example, can one determine the subject of a review when it is not explicitly named but only represented by the image of its logo?) Using XSLT creatively in XProc pipelines, the authors took even the barest of hints of semantic material to drive searching beyond the immediate HTML files for supplementary information. From this process they were able to construct a new, and consistent, collection of reviews with appropriate search tools.

The paper "Managing semantics in XML vocabularies: an experience in the legal and legislative domain," by a group at the University of Bologna led by Fabio Vitali, examined the structure of an information base sponsored by the United Nations, "Akoma Ntoso." (The name, in the Akan language of West Africa, represents understanding and agreement.) Akoma Ntoso collects parliamentary debates, laws, court rulings, and legal decisions needed to support open access to government processes. The problems faced by the developers included capturing the source documents in XML without altering their essential nature (since the content has legal implications) and erecting a strong metadata structure over the documents to allow searching and linking. In many cases the metadata must be quite extensive, for example, to identify the person speaking in a debate and why her participation is significant. Tagging of XML documents must be flexible, since none of the legal text can be obscured, and the system must be used by a wide variety of editorial staff in many countries. Some of the semantic layer is based on widely used archiving standards, such as FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records), and the expression of the layer is done so that multiple semantic tools (e.g., RDF/OWL or Topic Maps) could be applied for inferencing and navigation. This paper might have some relevancy to Y-12's problems with data archiving because its techniques address the need to store diverse information with minimally invasive markup, while providing a rich environment for retrieval.

Lynne Price (Text Structure Consulting) presented a paper that asked the question "DITA or Not?" DITA (Darwin Information Typing Architecture) is an XML based methodology developed by IBM to

support reusable building blocks for creating documentation. Through a survey, still in progress, Price has found that there are many misconceptions about DITA and how it should be used. Because many commercial XML systems (such as Arbortext, which is widely used in DOE) support DITA, naïve users are sometimes drawn into the system without understanding the underlying methodology. Far from obviating the need for document analysis and schema construction, as some have been led to believe, appropriate use of DITA requires extensive analysis and customization. (We have considered DITA for some Y-12 projects and concluded that it is frequently easier and quicker to adapt existing applications than to build *good* DITA applications.)

Walter E. Perry (Fiduciary Automation) often presents provocative ideas in his papers. With "IPSA RE: A New Model of Data/Document Management, Defined by Identity, Provenance, Structure, Aptitude, Revision and Events" he presented a layered structure that builds a hierarchy of financial transactions that can be transparent or opaque according to need-to-know criteria based on the legal notion of "privity." While Perry has not yet discussed details of how he implements this in XML, he has an intriguing notion that the XML objects he is constructing are verbs, not nouns (the usual view); several Topic Maps specialists find his model to be congruent with their thinking. Perry thinks he will implement his system on Google's App Engine API and BigTable. Perry's ideas might be particularly applicable to DOE's needs for building large but highly compartmented collections of secure data.

In "Where XForms Meets the Glass: Bridging Between Data and Interaction Design,"Andrew Spyker (IBM) presented some concepts that might be useful for Y-12 applications. There are many models for interaction and data communication in Web applications, particularly the AJAX (Asynchronous JavaScript and XML) model. The W3C XForms model can be integrated into AJAX processing to provide and end-to-end XML path for data that nonetheless provides interactivity and data validation in the browser while minimizing bandwidth requirements for connection with the server. Spyker showed several design patterns that are implemented using the Dojo Toolkit (a backing store for browser widgets).

One of the most important issues for maintaining XML documents in archives is the nature of document histories. Jean-Yves Vion-Dury (Xerox Research Centre Europe) presented "Stand-alone Encoding of Document History (or One Step Beyond XML Diff)," an update on a subject he had discussed last year. Vion-Dury presented a formal mathematical framework for expressing changes to an XML document, avoiding some of the known difficulties with what is commonly called "XML diff," while looking ahead towards implementation of the W3C project for "XQuery Update." Vyacheslav Zholudev (Jacobs University Bremen), also updating work presented last year, discussed "Scripting Documents with XQuery: Virtual Documents in TNTBase." The TNTBase system is a versioned XML-database with a client-server architecture, using Subversion and the Berkeley DB XML. Virtual documents are a proposed solution to the problem of creating an editable XML document from components stored in a content-management system. Zholudev's work shows how to locate documents, using an OWL/RDF framework, and enable the system online through a Web interface.

CONCLUSION AND RECOMMENDATIONS

DOE has been supporting the development of XML and related markup technologies for more than twenty-five years. The projects we are conducting under the PRIDE program continue to demonstrate that we are taking concrete steps to apply those technologies at the core of Y-12s manufacturing mission. The project we presented at *Balisage* in 2008 shows that we have taken advantage of XML technology in ways that (1) assure that we can make Y-12 product-certification data independent of any vendor's products, (2) we can encapsulate enough information in a data package to ensure that it will be usable for the projected stockpile life span of the products it documents, and (3) the certification data in the packages will be usable across DOE, NNSA, and the NWC, regardless of the site or the software used to read the XML data. Other PRIDE projects, and some begun on internal initiatives, similarly show a commitment to preserving data in nonproprietary formats and making it available for multiple processes.

This conference was very useful to us because it addressed many issues we will face in packaging manufacturing data in XML and maintaining it over very long time spans. Even technology intended to support fields that might not appear to be related to manufacturing, such as linguistics and literature, nonetheless can be adapted to solve problems facing the NWC. (Literary and historical researchers, for example, are concerned with preserving documentation of spans of centuries.) The papers heard and the discussions participated in brought us into contact with the latest thinking and best practices in XML technologies.

As Y-12 introduces increasing amounts of XML into its business practices, it needs to continue its involvement with core XML standardization. Because this conference is more closely aligned with the needs of Y-12, NNSA, and DOE than other XML conferences (which tend to focus on e-commerce), we should also continue to participate in it as a means of keeping abreast with current developments.

My participation in the Balisage 2010 conference was supported in part by the organizers of the conference.

This trip report was prepared in XML using the Arborext Editor with the application developed under the PRIDE project "Surveillance Reports Collaborative Authoring."

APPENDIX A

ITINERARY, 1–7 AUGUST 2010

Dates	Location	Contacts	Purpose
1 August 2010	Knoxville, Montréal		Travel
2 August 2010	Montréal	Michael Sperberg-McQueen, (Black Mesa Technologies, Chairman)	Conference: International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML
3–6 August 2010	Montréal	B. T. Usdin (Mulberry Technologies, Chairman)	Conference: Balisage 2010
7 August 2010	Montréal, Knoxville		Return travel

APPENDIX B

PRINCIPAL CONTACTS

The following is the attendance list from the Symposium and Conference.

Najeeb S. Andrabi, TIBCO Software Inc., nandrabi@tibco.com

Tom Angelopoulos, Epocrates, t0angel0@pacbell.net

Pete Aven, MarkLogic, Pete.Aven@marklogic.com

Christopher Ball, meta Heuristica, Christopher.Ball@metaHeuristica.com, Christopher.R.Ball@gmail.com

Marla Banks, Library of Congress, mban@loc.gov

Piotr Bański, University of Warsaw, bansp@o2.pl

Gioele Barabucci, Università di Bologna, barabucc@cs.unibo.it

Syd Bauman, Brown University, Syd_Bauman@Brown.edu

Jeff Beck, NIH/NLM/NCBI, beck@ncbi.nlm.nih.gov

Abraham Becker, NIH/NLM/NCBI, beckera@ncbi.nlm.nih.gov, abembecker@gmail.com

David J. Birnbaum, University of Pittsburgh, djbpitt@pitt.edu

Mario Blažević, Stilo Corporation, mblazevic@stilo.com

Naomi Bloch, Graduate School of Library and Information Science (GSLIS),University of Illinois at Urbana-Champaign, bloch2@illinois.edu

Eric Bloch, MarkLogic, Eric.Bloch@marklogic.com

Benjamin Bock, University of Leipzig, bock@informatik.uni-leipzig.de

Cyril Briquet, McMaster University, cyril.briquet@acm.org

Stephen Buxton, MarkLogic, Stephen.Buxton@marklogic.com

Kevin Caliendo, Loyola University Chicago, kcaliendo@luc.edu

William Candillon, 28msec Inc., william.candillon@28msec.com

Terry Catapano, Columbia Universit,y thc4@columbia.edu

Hugh A. Cayless, NYU, hugh.cayless@nyu.edu

Rebecca Clark, Aeon LLC, rclark@aeonxml.com

Florence Clavaud, École nationale des chartes (Paris, France), florence.clavaud@enc.sorbonne.fr, florence.clavaud@free.fr

Andrew Dombrowski, University of Chicago, adombrow@uchicago.edu

Quinn Dombrowski, University of Chicago, quinnd@uchicago.edu

Elizabeth H. Dow, School of Library and Information Science, Louisiana State University

Micah Dubinko, MarkLogic, Micah.Dubinko@marklogic.com

Iain Finlayson, MarkLogic, Iain.Finlayson@marklogic.com

Eric Freese, Aptara, eric.freese@aptaracorp.com

Tonya Gaylord, Mulberry Technologies, Inc, tgaylord@mulberrytech.com

Florent Georges, H2O Consulting, fgeorges@h2oconsulting.be, fgeorges@fgeorges.org
Renhart Gittens, Bodleian Libraries, Oxford University,rehart.gittens@bodleian.ox.ac.uk
Ian E. Gorman, Government of Canada, iegorman@gmail.com
Michael Greenlee, University of Illinois Urbana-Champaign, michael.greenlee@gmail.com
Cathy Moran Hajo, Margaret Sanger Papers Project, New York University, cathy.hajo@nyu.edu
Betty Harvey, ECC, Inc,. harvey@eccnet.com
Stefanie Haupt, Bielefeld University, st.haupt@gmail.com
Kevin S. Hawkins, University of Michigan Library, kevin.s.hawkins@ultraslavonic.info, kshawkins@fastmail.fm
Erik Hennum, efhennum@gmail.com
Mary Holstege, MarkLogic, mary.holstege@marklogic.com
Claus Huitfeldt, University of Bergen, Claus.Huitfeldt@uib.no
Matt Johnson, LexisNexis, matthew.c.johnson@lexisnexis.com
Michael Kay, Saxonica, mike@saxonica.com
William Eliot Kimber, Really Strategies, Inc,. ekimber@reallysi.com
Andrei Kolotev, NIH/NLM/NCBI, kolotev@ncbi.nlm.nih.gov
Anthony Laerdahl, The National Archives of Norway, tony@arkivvereket.no, anlar@arkivvereket.no
Matthew Landgraf, US Government Printing Office, mlandgraf@gpo.gov
Debbie Lapeyre, Mulberry Technologies, Inc., dalapeyre@mulberrytech.com
David A. Lee, Epocrates Inc., dlee@epocrates.com
Amelia A. Lewis, TIBCO Software, Inc,. alewis@tibco.com
Joshua Lubell, NIST, lubell@nist.gov
Yves Marcoux, Université de Montréal, yves.marcoux@umontreal.ca
Deborah Maron, Dumbarton Oaks (Harvard University,) deborah.maron@gmail.com
Fernando Mesa, MarkLogic, Fernando.Mesa@marklogic.com
John Meyer, ITHAKA/Portico, john.meyer@ithaka.org
Sheila M. Morrissey ITHAKA/Portico, sheila.morrissey@ithaka.org
Steve Newcomb, Coolheads Consulting, srn@coolheads.com
Dennis O'Connor, Cleared Solutions Inc., djo@clearedsolutions.com
Evan Owens, American Institute of Physics, eowens@aip.org
John Pedersen, John Wiley & Sons, jpederse@wiley.com
Simon Pepping, Elsevier, sampepping@gmail.com
Walter Perry, Fiduciary Automation, New York wperry@fiduciary.com
Terje Pettersen-Dahl, The National Archives of Norway, tepe@arkivvereket.no
Wendell Piez, Mulberry Technologies, Inc,. wapiez@mulberrytech.com
Denis Pondorf, University Bremen, Germany, pondorf@email.com
Luis Porras, Bombardier Aerospace, Luis.Porras@aero.bombardier.com, Luis.Porras@rogers.com

Pierre-Édouard Portier, LIRIS INSA-Lyon, pierre-edouard.portier@insa-lyon.fr
Lynne A. Price, Text Structure Consulting, Inc., lprice@txstruct.com
Martin Probst, EMC, Probst_Martin@emc.com
Liam R E Quin, W3C, liam@w3.org
Allen Renear, GSLIS, University of Illinois Urbana-Champaign, renear@illinois.edu
Hans-Jürgen Rennau, bits GmbH, rennau@bits-ac.com, hrennau@yahoo.de
Laine G.M. Ruus
Kenneth Sall, SAIC, kensall@gmail.com
Oliver Schonefeld, Institute for the German Language (IDS), schonefeld@ids-mannheim.de
Lisa Seaburg, Aeon LLC, lisa@aeonxml.com
Stephen Smith, Bombardier Aerospace, stephen.smith@aero.bombardier.com
Cecil Somerton, Fisheries and Oceans Canada, cecil.somerton@dfo-mpo.gc.ca
C. M. Sperberg-McQueen, Black Mesa Technologies, LLC, cmsmcq@blackmesatech.com
Andrew Spyker, IBM, aspyker@us.ibm.com
Mary Stephens, Bombardier, mary.stephens@aero.bombardier.com
Matthew Stoeffler, ITHAKA, matthew.stoeffler@ithaka.org
Maik Stührenberg, Bielefeld Universit,y maik.stuehrenberg@uni-bielefeld.de
Petter Svendsen, The National Archives of Norway, petter.svendsen@arkivvereket.no
Andreas Tai, Technische Universität München, andreas.tai@gmail.com
Marc Tessier, GOC, mrtessier@sympatico.ca
Vojtěch Toman, EMC, toman_vojtech@emc.com
Janine Trakhtenberg, JustSystems, janine.trakhtenberg@justsystems.com
Kimberly Tryka, NIH/NLM/NCBI, trykak@ncbi.nlm.nih.gov
B. Tommie Usdin, Mulberry Technologies, Inc., btusdin@mulberrytech.com
Jean-Yves Vion-Dury, Xerox Research Centre Europe, viondury@xeroxlabs.com, jy.viondury@gmail.com
Priscilla Walmsley, Datypic, pwalmsley@datypic.com
Norm Walsh, MarkLogic, Norm.Walsh@marklogic.com
Colleen Whitney, MarkLogic, Colleen.Whitney@marklogic.com
Karen Wickett, University of Illinois, wickett2@illinois.edu
David Williams, Woodward Governor Company, dawill@woodward.com
Sam Wilmott, sam@wilmott.ca
Sven Windisch, Topic Maps Lab, University of Leipzig, windisch@informatik.uni-leipzig.de
Andreas Witt, Institute for the German Language (IDS,) witt@ids-mannheim.de
Ann Wrightson, NHS Wales, a.m.wrightson@bcs.org.uk
Chris Zarate, Modern Language Association, czarate@mla.org
Wei Zhao, OCUL Scholars Portal, w.zhao@utoronto.ca
Vyacheslav Zholudev, Jacobs University Bremen, vyacheslav.zholudev@gmail.com
Kate Zwaard, US Government Printing Office, kzwaard@gpo.gov

APPENDIX C

LITERATURE ACQUIRED

The cumulative proceedings of the conference Balisage 2010 are available online at http://www.balisage.net/Proceedings/index.html.

Papers cited in the report include the following (at the time of writing not all the papers were available online*):

Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML

Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Balisage Series on Markup Technologies vol. 6 (2010), http://www.balisage.net/Proceedings/vol6/cover.html.

Beck, Jeff. "Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles," http://www.balisage.net/Proceedings/vol6/html/Beck01/BalisageVol6-Beck01.html, doi:10.4242/BalisageVol6.Beck01.

Dombrowski, Andrew, and Quinn Dombrowski. "A formal approach to XML semantics: implications for archive standards," http://www.balisage.net/Proceedings/vol6/html/Dombrowski01/BalisageVol6-Dombrowski01.html, doi:10.4242/BalisageVol6.Dombrowski01.

Lubell, Joshua. "Metadata for Long Term Preservation of Product Data," http://www.balisage.net/Proceedings/vol6/html/Lubell01/BalisageVol6-Lubell01.html, doi:10.4242/BalisageVol6.Lubell01.

Meyer, John, et al. "Portico: A Case Study in the Use of XML for the Long-Term Preservation of Digital Artifacts ," http://www.balisage.net/Proceedings/vol6/html/Morrissey01/BalisageVol6-Morrissey01.html, doi:10.4242/BalisageVol6.Morrissey01.

Quin, Liam R. E. "Beyond Eighteen Wheels: Considerations in Archiving Documents Represented Using the Extensible Markup Language," http://www.balisage.net/Proceedings/vol6/html/Quin01/BalisageVol6-Quin01.html, doi:10.4242/BalisageVol6.Quin01.

Schonefeld, Oliver, Andreas Witt, et al. "Sustainability of Linguistic Resources Revisited," http://www.balisage.net/Proceedings/vol6/html/Witt01/BalisageVol6-Witt01.html, doi:10.4242/BalisageVol6.Witt01.

Balisage 2010

Proceedings of Balisage: The Markup Conference 2010, Balisage Series on Markup Technologies vol. 5 (2010), http://www.balisage.net/Proceedings/vol5/cover.html.

---

*doi: Digital Object Identifier, a persistent online reference citation. See http://www.crossref.org/.

Cayless, Hugh A., and Adam Soroka. “On Implementing string-range() for TEI,” http://www.balisage.net/Proceedings/vol5/html/Cayless01/BalisageVol5-Cayless01.html, doi:10.4242/BalisageVol5.Cayless01.

Freese, Eric. “Multi-channel eBook production as a function of diverse target device capabilities,” http://www.balisage.net/Proceedings/vol5/html/Freese01/BalisageVol5-Freese01.html, doi:10.4242/BalisageVol5.Freese01.

Georges, Florent. “The EXPath Packaging System: A framework to package libraries and applications for core XML technologies,” http://www.balisage.net/Proceedings/vol5/html/Georges01/BalisageVol5-Georges01.html, doi:10.4242/BalisageVol5.Georges01.

Haupt, Stefanie, and Maik Stührenberg. “Automatic upconversion using XSLT 2.0 and XProc: A real world example,” http://www.balisage.net/Proceedings/vol5/html/Haupt01/BalisageVol5-Haupt01.html, doi:10.4242/BalisageVol5.Haupt01.

Huitfeldt, Claus, Yves Marcoux, and C. M. Sperberg-McQueen. “Extension of the type/token distinction to document structure,” http://www.balisage.net/Proceedings/vol5/html/Huitfeldt01/BalisageVol5-Huitfeldt01.html, doi:10.4242/BalisageVol5.Huitfeldt01.

Kay, Michael. “A Streaming XSLT Processor,” http://www.balisage.net/Proceedings/vol5/html/Kay01/BalisageVol5-Kay01.html, doi:10.4242/BalisageVol5.Kay01.

Lewis, Amelia A, and Eric E. Johnson. “gXML, a New Approach to Cultivating XML Trees in Java,” http://www.balisage.net/Proceedings/vol5/html/Lewis01/BalisageVol5-Lewis01.html, doi:10.4242/BalisageVol5.Lewis01.

Perry, Walter E. “IPSA RE: A New Model of Data/Document Management, Defined by Identity, Provenance, Structure, Aptitude, Revision and Events,” http://www.balisage.net/Proceedings/vol5/html/Perry01/BalisageVol5-Perry01.html, doi:10.4242/BalisageVol5.Perry01.

Portier, Pierre-Édouard, and Sylvie Calabretto. “Multi-structured documents and the emergence of annotations vocabularies,” http://www.balisage.net/Proceedings/vol5/html/Portier01/BalisageVol5-Portier01.html, doi:10.4242/BalisageVol5.Portier01.

Price, Lynne A. “DITA or Not?,” http://www.balisage.net/Proceedings/vol5/html/Price01/BalisageVol5-Price01.html, doi:10.4242/BalisageVol5.Price01.

Renear, Allen H., and Karen M. Wickett. “There are No Documents,” http://www.balisage.net/Proceedings/vol5/html/Renear01/BalisageVol5-Renear01.html, doi:10.4242/BalisageVol5.Renear01.

Rennau, Hans-Jürgen. “Java Integration of XQuery - an Information-Unit Oriented Approach,” http://www.balisage.net/Proceedings/vol5/html/Rennau01/BalisageVol5-Rennau01.html, doi:10.4242/BalisageVol5.Rennau01.

Stührenberg, Maik, and Christian Wurm. “Refining the Taxonomy of XML Schema Languages. A new Approach for Categorizing XML Schema Languages in Terms of Processing Complexity,” http://www.balisage.net/Proceedings/vol5/html/Stuhrenberg01/BalisageVol5-Stuhrenberg01.html, doi:10.4242/BalisageVol5.Stuhrenberg01.

Vion-Dury, Jean-Yves. “Stand-alone Encoding of Document History (or One Step Beyond XML Diff),” http://www.balisage.net/Proceedings/vol5/html/Vion-Dury01/BalisageVol5-Vion-Dury01.html, doi:10.4242/BalisageVol5.Vion-Dury01.

Wickett, Karen M. “Discourse situations and markup interoperability,” http://www.balisage.net/Proceedings/vol5/html/Wickett01/BalisageVol5-Wickett01.html, doi:10.4242/BalisageVol5.Wickett01.

Zholudev, Vyacheslav, and Michael Kohlhase. “Scripting Documents with XQuery: Virtual Documents in TNTBase,” http://www.balisage.net/Proceedings/vol5/html/Zholudev01/BalisageVol5-Zholudev01.html, doi:10.4242/BalisageVol5.Zholudev01.

DISTRIBUTION

DOE DISTRIBUTION

Ms. C. S. Blackston, U. S. Department of Energy, 19901 Germantown Road, IM-21, Room C-137, Germantown, MD 20874-1290

Mr. David Bellis, Office of Scientific and Technical Information, OSTI

Mr. Philip A. Carpenter, DOE-ORO ORNL Site Office

Ms. Debbie Cutler, OSTI, P.O. Box 62, Oak Ridge, TN 37831

Ms. Pamela. L. Gorman, DOE-ORO Y-12 Site Office

Ms. Kelli Holden, Bldg. K-1030, MS-7312

Mr. Paul F. Krumpe, U. S. Department of Energy, NA-122.21, Forrestal Building, Washington, DC 20585

Mr. Bruce E. Lownsbery, Lawrence Livermore National Laboratory Mailcode L-170 7000 East Ave., P.O. Box 808 Livermore, CA 94550

Mr. Donat R. St. Pierre, Safeguards and Security, ORO

Mr. Lawrence Sanchez, U. S Department of Energy, IN-1, Room GA-301, Forrestal Building, Washington, DC 20585

Mr. Theodore D. Sherry, DOE-ORO Y-12 Site Office

Mr. Robert Shoup, Decision Applications, Division D-6 Risk Analysis & Decision Support, Los Alamos National Laboratory, Mail Stop A120, Los Alamos, NM 87545

Dr. Andrew P Weston-Dawkes, U. S. Department of Energy, HS-90, 19901 Germantown Road, Germantown, MD 20874-1290

Mr. B. R. White, U. S. Department of Energy, 19901 Germantown Road, IM-60.2, Room C-137, Germantown, MD 20874-1290

Threat Reduction Team, 4x24 NHB, Washington, DC 20505

Office of Scientific and Technical Information, OSTI

INTERNAL DISTRIBUTION

E. E. Angros

D. R. Baumgardner

R. Baylor

C. A. Barton

M. A. Bell

V. E. Chase

K. M. Catlett

G. A. Dailey

T. C. Domm

J. T. Fisher

D. R. Hamrin

S. E. Hughes

T. M. Insalaco

D. M. Kelleher

J. S. Kyle

K. E. Langley

D. J. Linehan

R. L. Luttrell

M. A. McNeil

J. D. Mason

P. E. Parris

B. K. Robinette

R. C. Secrist

R. L. Shipp

K. J. Smith

T. O. Tallant

R. M. Wilson

Y-12 Plant Records Department

Foreign Travel Office—RC

EXTERNAL DISTRIBUTION

Ms. Barbara Beeton, American Mathematical Society, 201 Charles Street, P.O. Box 6248, Providence, RI U.S.A. 02940

Dr. Doris Bernardini, DISA Interoperability Directorate, 20047 Presidents Cup Terrace, Ashburn VA 20147-4119

Dr. Michel Biezunski, 402 85th Street, No. 5, Brooklyn, NY 11209

Mr. Jon Bosak, 1448 Trumansburg Road, Ithaca, NY 14850

Mr. Martin Bryan, IS-Thought, 29, Oldbury Orchard, Churchdown, Gloucester GL3 2PU, United Kingdom

Mr. Francis Cave, Francis Cave Digital Publishing, The Old Bakery, Felday Glade, Holmbury St Mary, Dorking, Surrey RH5 6PG, United Kingdom

Mr. Robin Cover, Isogen International, 6634 Sarah Drive, Dallas, TX USA 75236

Mr. James Bryce Clark, OASIS, 630 Boston Road, M102, Billerica, Massachusetts 01821

Mr. Patrick Durusau, 2149 Conyers St SE, Covington, GA 30014

Dr. Martin J. Fritts, 1710 SAIC Drive, P.O. Box 1303, Mail Stop 2-6-9, McLean, VA 22102

Dr. Charles F. Goldfarb, Information Management Consulting, 13075 Paramount Drive, Saratoga, CA 95070

Mr. G. Ken Holman, Crane Softwrights Ltd., 1605 Mardick Court, Box 266, Kars, Ontario K0A-2E0, Canada

Dr. Richard L Klobuchar, SAIC, 2829 Guardian Lane, Virginia Beach, Va 23452

Dr. Yushi Komachi, Osaka Institute of Technology, Faculty of Information Science, and Technology, 1-79-1 Kitayama, Hirakata, Osaka, 573-0196, Japan

Dr. Peter Kortman, 2109 St. Ives Blvd., Knoxville, TN 37922

Ms. Mary McRae, OASIS, 630 Boston Road, M102, Billerica, Massachusetts 01821

Dr. Steven R. Newcomb, 268 Bonnet Way SE, Southport, NC 28461 USA

Dr. Sam Gyun Oh, Sungkyunkwan University, Myungryun-Dong 3 Ga 53, Chongro-Gu, Seoul 110-745 Korea

Mr. Charles Onstott, SAIC, 301 Laboratory Road, P.O. Box 2501, Oak Ridge, TN, 37831

Mr. Steve Pepper, Maridalsveien 99b, N-0461 Oslo, Norway

Dr. Lynne Price, Text Structure Consulting, Inc., 17225 San Franciscan Drive, Castro Valley, CA 94552

Mr. Norman Smith, SAIC, 301 Laboratory Road, P.O. Box 2501, Oak Ridge, TN, 37831

Dr. Richard Strehlow, Gourmet's Market, 5107 Kingston Pike, Knoxville, TN 37919

Ms. B. Tommie Usdin, Mulberry Technologies, Inc., 17 West Jefferson Street, Suite 207, Rockville, MD 20850

Dr. Charles Wilson, Areteq, Inc., 132 County Road 884, Etowah, TN 37331

Ms. Ann Wrightson, Informing Healthcare (NHS Wales), 10-11 Old Field Road, Bocam Park, Pencoed, Bridgend CF35 5LJ, United Kingdom