Report on the ERPANET Workshop on XML for Digital Preservation, Urbino, 9 – 11 October 2002[1]

Maureen Potter
Experiment Operator
Digital Preservation Testbed
5 November 2002

Summary

The workshop was, overall, a great success. Some of the sessions had more to offer in terms of practical experiences than others, but both audience and speakers recognised this and the discussions that followed each session proved to be lively and relevant.

Schedule:

Five sessions were scheduled, for the morning and afternoon of each day and the morning only of the Friday. Each session consisted of two or three speakers. A 'breakout group' took place at the end of each session, to discuss aspects of the presentations, ask questions of the speakers, and try to relate the session back to research questions structured by Erpanet.

Day One: Morning
Session title: Introduction to XML as a Digital Preservation Strategy.

Chair: Maria Guercio (Italian Erpanet Director)

After an introduction and welcome by Maria Guercio, the first session was devoted to a technical presentation on the various aspects of XML - XSLT, CSS, FO, SGML-HTML relationships etc. This was presented by Giovanni Michetti of the University of Urbino and was titled 'Principles and Potential Role of XML. The audience was of mixed ability, so this was useful for those who were of mid-level, but, as one attendee later testified, slightly overwhelming for those just beginning on the XML path. Giovanni mentioned that they have been working with the Ministry of Finance to analyse electronic documents and look for a DTD that will cover them all; if there is not one that will cover them all then they will design different ones for different purposes.

The breakout group that followed was probably the least fruitful of the five, because the presentation had focused on technical issues and the questions scheduled for the session covered a far broader range of issues. However, a resounding point to come out of the session was the reminder that we are all trying to preserve slightly different things, and that preservers must define what it is they are trying to preserve. Giovanni believed that boundaries between different types of digital objects are becoming blurred, but others disagreed. He also believed that we must get involved as soon as possible – 'working upstream to preserve the stuff that later comes downstream'. Andrew Wilson of the National Archives of Australia also gave a brief outline of their digital preservation research and mentioned that they will be investigating converting spreadsheets to XML.

---

[1] This report was originally intended as an internal document to communicate our learning experiences from Urbino with the rest of the Testbed team. It has been constructed through notes and memory; apologies to anyone whose presentation I may have misrepresented.

Day One: Afternoon
Session Title: XML as a Standard for Metadata Preservation and Metadata Exchange.
Chair: Hans Hofman (Netherlands Erpanet Director)
Stephan Heuscher of the Swiss Federal Archives presented two case studies: the first on SIARD, Software Independent Archiving of Databases; and the second on AMDA, Audio Metadata Acquisition. He opined that documentation for the XML files was extremely important, and said that XML was not as widely supported in software tools and by manufacturers as is widely believed. He described the SIARD case study in depth, stating that SQL99 was used for the low-level data description – so the primary data was not in XML but used XML as a wrapper.

The second presentation, by Carlo Batini, provided a case study on the legislation of Italian current records. This was a highly promising presentation but the speaker had to rush off afterwards. He spoke of their work with email, saying that registration numbers were added to all emails to get them into the workflow system, and that they were obliged to add DTD's to some records and mark them up in XML (although I did not catch which records or why). They have three types of DTD:

Basic DTD:	For documents with regular structure that respected drafting rules
Strict DTD:	For documents with complex structure that respected drafting rules
Loose DTD:	For documents with irregular structure, exceptions, and documents not respecting drafting rules [Useful for old historical documents].

They are producing a tool to work with Word (some sort of macro) that will mark up the documents in XML.
This was definitely one of the more relevant papers, but there was no relation of specific details. Contact: batini@aipi.it

Enrico Rendina of the Consorzio di Roma Ricerche, who spoke about XML experience for Historical Archives, presented the final paper of the session. The language was complex - for example, 'Backward Retrieval' was defined as the transposition of pre-existing information into a different medium, but that sounds like migration (One or two other speakers suffered from this complexity-affliction too). Enrico spoke of the scanning and conversion of documents into XML, but the document itself wasn't marked-up - only the metadata. DTD's were established and conformed to EAD (Encoded Archival Description) standards and requirements.

The breakout session discussed the use of XML as a tool, as opposed to a strategy. Most of the technicians in the room spoke of XML as a tool, NOT a strategy, but the archivists and others were more amenable to using XML as part of a wider approach and viewed it as more than just a tool. Problematic creation practices were discussed, and we returned to this point in the session the following day. Again, we discussed what we were trying to preserve, and asked 'what is the problem we are trying to solve?' This varies for different people. For Testbed, it is interoperability AND digital obsolescence, both of which XML can address. For others, it was merely interoperability; for others again, it was a requirement for a clear and structured way of marking data. XML obviously meant more to some people than to others.

Day Two: Morning
Session title: XML as a Preservation Method

Chair: Seamus Ross (UK Erpanet Director)

Fynette Eaton of NARA opened the session with a presentation called 'NARA Explores Possible Uses of XML'. This was a largely general paper, outlining the challenges NARA are facing and the direction they may take in the future. The Persistent Object Preservation Approach was mentioned, which involves migrating objects out of their original format. Fynette outlined NARA's interest in the following topics:

- Using XML DTD's or XML Schema to manage attributes and semantics
- Using XML based tools to mediate between heterogeneous collections
- Using XTM (Topic Maps) DTD's to manage relationships
- Using XSL (Style sheets) to manage presentation

She closed by outlining future work, which included system design requirements and emails with attachments. Copies of the slides she presented should soon be available from the Erpanet website.

The second paper was presented jointly by Marco Rendina (Consorzio Roma Ricerche) and Salvatore Costa (Ministry of Finance) and discussed some of the work the Italian Ministry for Finance has carried out with the San Diego Super Computer Centre. This was a largely technical demonstration in which they showed how they migrated from EBCDIC and COBOL to ASCII, then to XML. The COBOL declarations were accompanied by comments inserted by the original file creator and it was a hierarchical structure. The XML tags were the original COBOL tags, and as the original COBOL file was well structured, the migration seemed to appear relatively straightforward. Additional documentation was required to describe the XML tagging. They used a sample from the Income Tax Returns, which contained approximately 50,000 records. The experiment took one week to run in San Diego and homogenous access was provided through a web interface.

Marco related the problems they experienced with the migration tools supplied, in that they 'simply didn't work'. As far as the topic of their preservation activity was concerned, they had two interests: the preservation of data (the digital object as a bit stream) and the preservation of information (the tagged data). He acknowledged that they still needed to work on logical-, functional/algorithmic-, and procedural- knowledge relationships.

We then presented the Testbed paper on XML for Archival Preservation. I introduced the Testbed, its background and scope, and some of our research questions. I then discussed the advantages and disadvantages of using XML for archival preservation, and talked about why XML was particularly suitable for emails. The bulk of the presentation focused on the three implementation options we have developed for conversion/provision of emails in XML. I outlined each of the approaches [termed the All-In-One Option (our Process 2), the Split Files Option (our Process 3), and the Forward Facing Option (our demonstrator)] and closed with a demonstration of the email to XML demo. The presentation was very well received.

The breakout session that followed involved the recall of all speakers in that session to the stage. We were asked about documentation for the XML files ('we don't need much for the emails; it's something we'll take into account with the more complex record types') and how we could get Users to use our add-in properly. This led to a discussion on record creation practices, the future direction of Microsoft, and the differences between the various approaches we have all developed.

Day Two: Afternoon
Session title: XML. Digital Preservation and the Software Market
Chair: Maria Guercio (Italian Erpanet Director)
The afternoon session used the concept of a Round Table with three Italian software companies: *Software AG-Tamino*; *Filenet*; and *3D.* Each company gave a presentation (some more sales focused than others), introduced their products, and provided some experience of their customers digital preservation needs. The software companies outlined their requirements for the meeting – namely, they wanted to know our requirements for digital preservation and for those requirements to be ideally embedded in law – preferably something along the lines of 'digital records should be preserved using XML'. Of course, this was far too simplistic an approach to work, and as the afternoon wore on, it became increasingly clear that the audience and the Round Table members were talking along very different lines. It was encouraging to see their presence at this meeting, but it was clear that much more work needs to be done, not only to define requirements for the different communities with an interest in digital preservation, but also to effectively communicate these requirements and different interests to the software market.

Day Three: Morning
Session title: XML and Web Archiving
Chair: Niklaus Bütikofer (Swiss Erpanet Director)
The final session of the workshop did not wholly conform to the title. Whilst Andreas Rauber (ERCIM Research Fellow at INRIA, France) spoke about web archiving in a presentation entitled 'Towards A European Web Archive', Regan Moore of the San Diego Supercomputer Centre spoke on Persistent Objects, with no reference to web archiving at all.

Andreas described the beginnings of a new project, for which calls for proposals will be issued at the end of the year. He discussed the various aspects of the project that need to be considered and defined. These included Acquisition of Information – source and data selection; Bit Preservation, including storage media migration and refreshing; and Logical Preservation, using system emulation and format compression. Strategies to investigate will include emulation, conversion, and abstraction. He raised the question of what should be preserved and how? The original? Its appearance or just plain content? Functionality? He noted that web pages require additional metadata such as frequency of updates, and contact information not only for the client of the website but also the designer and maintainer. He asked how far XML could take us in the preservation of web pages – can it represent applets, for example[2]? His presentation was interesting and raised some good issues, but focused largely on potential work, mentioning only briefly the results and strategies of web-archiving projects that have already been completed.

Regan Moore then spoke on 'Persistent Objects'. He first defined the goal of the Persistent Archive project as 'to identify the key technology that will facilitate the creation of a persistent archive of archival objects'. This requires the definition of a set of mechanisms to allow the management of various types of digital objects. He defined the following archival processes: appraisal; accession; arrangement; description; preservation; and access.
He discussed migration and emulation. He defined one of the problems of emulation as the use-restriction it places on the new user; they can only do that which the original user could do. Because of this, the SDCC prefer to migrate, allowing the user to manipulate a digital/record object with new and current technology. This use-case argument for their

[2] Note that the National Library of Australia have some experience with preserving Applets – see Colin Webb, *Digital Preservation – A Many Layered thing: Experience at the National Library of Australia* http://www.clir.org/pubs/reports/pub107/webb.html

choice of strategy also applies to technology management, especially access and display technology. Unfortunately, he did not provide in-depth details of their migration approach. He mentioned the process of Word to XHTML to XML, but did not explain why this path was chosen, nor why they went to XHTML instead of a pure XML conversion from which many representations could be produced. He also discussed the use of the Storage Broker Resource Element and Data Grids, Process Maps and Workflow.

The Breakout session that followed concentrated first on the web archiving presentation, then on the SDCC presentation. The topic of authenticity came up during the SDCC Q&A session. Regan was asked if the SDCC had defined authenticity requirements for their migrations, and if they had defined what is an 'acceptable level of change'. He answered that they haven't, although they are trying to define an accessioning template for different kinds of digital objects. This sparked a lively discussion on the requirements for preserving attributes, with Andrew (NAA) commenting that he thought look and feel were actually pretty irrelevant, and Regan adding that preservation of the intellectual capital was more pressing. Andreas commented that original look and feel could actually tell you a lot about content and context, and cited poetry and archaeology as an example. Overall, opinions were mixed, and no real conclusion was reached on the matter – hardly surprising as we are all at different stages of different investigations on the preservation of different types of digital objects with differing requirements!

The summary and conclusions for the workshop drew on the opinions and contributions made over the past few days. Seamus spoke on the need for processes and behavioral interests, many of which we currently had no way to represent. He also recognized the need for more (standardized?) development of parsers and viewing tools. We still have a long way to go in confidently using XML for preservation, but different groups are now considering many of the required aspects globally. It is very encouraging to see that the first steps have already been taken not only by the Testbed but also by other research groups, and countries such as Switzerland and New Zealand. There is practical experience out there, we must just look past the superpowers and focus on those smaller countries actively implementing and experimenting with approaches.

Maureen Potter
Experiment Operator
Digital Preservation Testbed
5 November 2002