

ERPANET WORKSHOP

XML as a Digital Preservation Strategy Urbino, 9-11 October 2002

Background paper

Introduction

XML is increasingly being used as a means for providing access and interoperability to digital information and for preservation purposes. The focus of this second ERPANET workshop will concentrate on using XML as a tool in preserving the long-term value of this digital information.

The capability of the eXtensible Markup Language for building and maintaining complex logical frameworks through the definition of tags and related grammars and semantics is a concrete opportunity to deal with the difficulties of the preservation process in different domains of the digital environment (web-archiving, digital libraries, museums collection metadata, electronic records and archival finding aids, etc.). Various methods are under evaluation, with reference to the costs involved, the feasibility and the quality and reliability of the results. XML approach looks today a promising tool for long-term preservation of digital object, able to be applied to collections as well as to individual objects. By the way, its effective and correct use implies the definition of schemas for different digital objects, both during the digital objects creation process and at the migration/preservation phases. Complex is also to guarantee the maintenance of the metadata relationships as essential information to preserve on time the logical structure of the records or the collections sets and aggregations.

The experience done till now should be further analysed and evaluated to verify the efficiency (and the difficulties) of the solutions in different areas of application, specifically when the digital objects typologies cannot be easily described by using uniform models. More than investigating the XML potentialities in this area (confirmed by the increasing interest of the cultural institutions but also by the public administrations and the business sector), the workshop will present and critically discuss the main projects developed in this area and focus on the crucial aspects of this experience to design advanced uses of the language and to identify issues that should be further explored to transform a potentiality into effective instruments and methods for long-term digital preservation.

The workshop will focus on five main areas:

- introduction in digital preservation (some principles) and the possible/potential role of XML
- the role of XML as standard for metadata preservation and metadata exchange
- XML as a preservation method
- XML, digital preservation and the software market

- XML and web archiving

Questions for discussion

Introduction in digital preservation and the possible/potential role of XML

1. What are the options for technical solutions to the preservation of archival records over the long term?
2. What are the advantages and disadvantages of XML for archival preservation?
3. What makes XML a particularly suitable method for specific types of digital objects like emails?
4. What is crucial in defining XML schemas for preserving various forms of digital objects?
5. Is it useful to establish partnerships in the development and use of XML and what is required?
6. What are the barriers that will prevent agencies and institutions from using XML?

The role of XML as standard for metadata preservation and metadata exchange

1. What are the issues re: metadata that can or cannot be solved with XML?
2. •What criteria for use can be identified?
3. •Exchange standard: what is necessary to make that effective? •What alternatives exist?Do we need different standards for metadata exchange and metadata preservation and why?
5. How should the XML structure be defined (DTD or Schema) and why?

XML as a preservation method

This session will require a different set of questions related to the various experience and research projects presented. Here you will find a first list of some basic questions related to the use of data grids to implement persistent archives (SDSC project, the Italian Ministry for finance experiment, NARA experience) and to the technical problems referred to the XML conversion of emails realized by the Dutch Digital Preservation Testbed Project

1. What capabilities are needed to automate the processes of appraisal, accession, arrangement, description, preservation, access, and re-purposing?
2. Is preservation made easier by the use of data grid technology? How help could XML provide
3. Can technology evolution be managed through use of data grid technology?
4. How can the information and knowledge context be defined for archival fonds or other kind of digital collections?
5. Can information and knowledge be managed using data grid technology?
6. What makes XML a particularly suitable strategy for a specific document typology like emails? What are the implementation options for providing/converting emails in(to) XML? How can we keep the different component parts of the message, including its metadata and attachments, inviolably linked?

XML, digital preservation and the software market

The round table is organised with the participation of software houses differently involved into the development of XML tools for digital objects management and preservation: the developers of a specific XML tool (Tamino developed by Software AG), a multinational company strongly involved in building EDMS and ERMS (Filenet) and an Italian company involved in the EDMS/ERMS which pays a special attention to the XML potentialities (3D Informatica). The speakers will be asked to discuss the reactions of the IT market to the use of XML in the past years and the future perspective with specific reference to the quality, the costs and the feasibility of long term digital preservation.

XML and web archiving

This session is only partially dedicated to XML implementation. Web archiving is becoming a crucial problem for the development of Information Society and its memory preservation. Its complexity is still unexplored and requires first the identification of the main basic questions:

1. what should be archived? what will be of interest to the future? do we need everything? all the time? is "half of it" sufficient? which applications can we address? which ones do we want to address?
2. what can we archive with respect to data acquisition, amount of data, technology? all the web? deep web? meta-information on the data, their producers, used technology?
3. what/how should it be preserved? preserve "original"? preserve appearance? preserve content? preserve functionality?
4. which strategies allow us to reach these goals? Is there "a strategy"? emulation? conversion (migration)? abstraction/extraction?
5. In how far can meta-data/XML help us to reach these goals, how far can it go from being merely a means for preservation to actually preserving the information and characteristics we want to preserve?

Abstracts

Carlo Batini (Autorità per l'informatica nella pubblica amministrazione): XML for the documents and records creation

The experience of *Norme in rete* (a strong effort already developed to create in XML all the policy records and documents and the regulations and legislation) and the discussion on the national policy which implies the use of XML to exchange any kind of records identifiers (registries and classification metadata, etc.)

Fynnette Eaton (US National Archives and Records Administration), NARA explores possible uses of XML

The speaker will provide information about the U.S. National Archives Electronic Records Archives Program for digital preservation, how it has developed and why it is considering the use of XML as part of the method for preserving access to electronic records over time. She will also discuss a current initiative in which the National Archives will ask Federal agencies to transfer specific types of electronic records with XML schemas.

Stephan Heuscher (Swiss Federal Archives)

Softening the borderlines of archives via XML - a case study (multi-source meta data acquisition of digital audio recordings of the sessions of the Swiss parliament) Archives have always had troubles getting metadata in formats they can process. With XML, these problems are lessening. Many applications today have the option of exporting data into an application-defined XML format that

can easily be processed (XSLT, schema mapper, etc) fit the archives' needs. In a practical example we have acquired existing metadata describing debates at the Swiss parliament from a legacy system as well as from an Augias (-Access) database. The internal storage of the AMDA (Audio MetaData Acquisition) is handled by a database. The export of the collected metadata will also be in XML into our digital archive (to be built). To end I'd like to talk about other experiences in the ARELDA project, especially the archiving of relational databases and the experiences with XML in this project. Specially to show our strategy: metadata in well-documented XML, primary data in an open standard format.

Maureen Potter, Digital Preservation Testbed Project

The purpose of the report is to illustrate how relatively simple extensions to standard software can go a long way to addressing the most important problems that organisations are experiencing with managing e-mail. It could also help in developing recommendations on preservation approaches. The project allows for two types of e-mail message: informal or formal. Informal messages are those which are not related to the official business of the department and are not archived. The only restriction which the e-mail tool applies to an informal e-mail is to insert a disclaimer at the top of the message to tell the recipient that this is not an official message. Formal messages are assumed to have the same status as a letter on the department's headed paper and we have used that analogy in several aspects of our approach. In a formal e-mail, the user is prompted to fill in a number of metadata items (Dossier, Programma, Handeling etc.) Other items, such as the user's name, organisation and contact details only need to be entered once and thereafter are filled in automatically. Outlook interacts with the user's Address Book or Contacts folder to extract additional information about the recipients of the message, which are also included in the message metadata. The customised version of Outlook combines the metadata and the message content into an XML file. This XML file is then transmitted to a central server, which checks the XML against a schema and stores the XML file in the archive. The server also applies an XSL (XML Stylesheet Language) stylesheet to the XML file to produce a formatted HTML file in the house style of the organisation, applying fonts, colours and logos for example. This HTML format message is sent back to Outlook to become the actual e-mail message which is sent to the recipients. In this way, the application can ensure that the e-mail message is archived together with the required metadata and at the same time, can apply a uniform house style to the official messages produced within an organisation.

Andreas Rauber, Towards a European Web Archive: Issues and Next Steps

With the growing importance of the Web and its evolution from a technological playground to one of the core infrastructures, an "information mega-store" with tremendous diversity of information artefacts, awareness has risen for the pressing need to archive it as an entity, i.e. the documents, their structure and technology, as well as to use the information constituted by it. Numerous initiatives thus are being created, aiming at the collection of on-line publications, databases, or Web pages in general, be it delivery or deposit of documents, selectional or free harvesting, their preservation for the future, or analysis of the Web in terms of content, structure, and technology.

The ambitious goal of a European Web Archive not only provides an indispensable asset for our digital cultural heritage. It also represents a mirror of society and its needs, communities and their languages, of technology and market evolution, with far-reaching consequences for numerous application domains, such as administration (e-government, e-democracy). In this talk, the initiative for the creation of a European Web Archive within the framework of an EU IST 6th Framework Integrated Project is presented, and questions pertaining to its goals and means to achieve them will be discussed.

Enrico Rendina (Consorzio Roma Ricerche), XML experience for historical archives

The presentation will discuss the use of XML for descriptive metadata management, specifically in the archival sector and with reference to a concrete experience applied to the automatic tracking of existing findings aids metadata. The aim is to stress the potentialities of XML technologies (X-Path, XSLT, XSL-FO etc.) by using metaphoric examples.