IST-2001-32706

Urbino Workshop Report

9-11 October 2002

DOCUMENT NAME:	Urbino Workshop: XML FOR DIGITAL PRESERVATION
DATE:	22.11.02
WORK PACKAGE:	WP 5
TASK NUMBER:	5.1
AUTHOR TITLE:	Italian Content Editor
CONTACT DETAILS:	Italian.Editor@erpanet.org

VERSION: Draft ■ Final □

ERPANET WORKSHOP
XML FOR DIGITAL PRESERVATION. FINAL REPORT

Urbino, 9-11 October 2002

Executive Summary

ERPANET held its second experts workshop in October in Urbino. The event offered experts the opportunity to take part in an investigation of key topics in digital preservation. Participants examined XML role in digital preservation. This report is the result of two and a half days of presentations and discussions, and it aims to enable practitioners, their institutions and funding agencies to improve conservation and preservation practices.

Speakers came from all over Europe and from diverse fields. They introduced key topics to the expert audience and this report is a blend of the talks and discussions that took place. The report makes some conclusions with reference to the potentials of XML and the risks involved in the lack of awareness about the requirements of the digital resources to be preserved:

- XML makes the automatic acquisition of metadata possible if the creators develop advanced analyses and tools of knowledge management and structures representation,
- This advanced use concerns the capability and need to describe and maintain data, information and the knowledge connected to the documentary resources: more experimentation is required,
- Need to encourage the use of XML in the most advanced and complete forms, i.e. planning DTD and descriptive models of the different types of digital resources to be preserved,
- Importance of documentation relating to the creation of digital sources – including the description of behaviours and processes- in view of the difficult and expensive retroactive retrieval of new records,
- Availability of poor processes to train and make aware the creators on the representation tools for digital objects, mainly in the case of complex resources,
- Need to change the organizational approach of digital objects and documents producers in order to ensure early and effective actions including cost-benefit analyses.

Introduction

The second workshop organised by the Università degli Studi di Urbino, in the context of the ERPANET European project, was devoted to the use of XML for digital preservation. 62 participants from many Italian and international institutions involved in the creation and preservation of digital resources (museums, archives, libraries) were present. The proceedings followed the timetable indicated in the programme: each session included two or three presentations followed by lively debates reported in this paper. The programme included a visit to the Ducal Palace and a dinner offered by the University of Urbino. In fact the University supported the initiative through the research funds of the COFIN 2000 project on long-term preservation of electronic documents.

The use of XML as a tool to guarantee interoperability among current systems and long time preservation has been the focus of many international research projects and it deserves special attention. The Urbino workshop has been the first opportunity to discuss the facilities of the tool and assess the experiences obtained, mainly in the public sector, as mentioned in the background paper and underlined by Mariella Guercio in presenting the programme of the initiative:

> "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 object typologies cannot be easily described by using uniform models. More than investigating XML (confirmed by the increasing interest of cultural institutions but also by 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 outline advanced uses of the languages and identify the issues that should be further explored to transform potentials into effective instruments and methods for long-term digital preservation".

Aims and Objectives of the Workshop

Over the course of the three days, participants were provided with papers and examples presented by experts before analysing and discussing the issues and challenges arising from the use of XML in digital preservation.

As clearly expressed in the project funded by European Commission, ERPANET workshops are not directed to a general technical audience, but more specifically to the preservers, institutions and professionals already involved in the technical problems (from different points of view) of creating and preserving digital objects. Consequently, the workshop of Urbino has been organised with the aim of offering a framework able to promote discussion on specific topics: the organizers have prepared a first list of questions for each session to help the audience in defining and debating the crucial aspects of the workshop subject. Each session was introduced by at least two presentations, followed by a long discussion aiming not only at clarifying and redefining the issues partially listed in the introductory document but also identifying recommendations and remarks useful to restrict the main topics of the workshop and provide operational indications to support the institutions and operators involved.

The workshop has been organised into six sessions:

- Introduction on digital preservation and the potential role of XML
- The role of XML as a standard for metadata preservation and metadata exchange
- XML as a preservation method (critical analysis of case studies)
- XML, digital preservation and the software market (a round table)

- XML and web archiving (European initiatives)
- Conclusions

For each session the questions listed in the background paper are the synthesis of the suggestions made by the speakers and the organisers. In some cases the questions have been too many and too complex to be answered in the short time of a session but they have been able to orient the discussion and provide a general framework for further investigations.

This report is organised into six sections, with the first five dealing with individual workshop sessions and the final section presenting the recommendations and conclusions on the agreements reached by the participants.

Session 1: Introduction on digital preservation and the potential role of XML

The first session has been characterised by a methodological introduction, presented by Giovanni Michetti, on XML as a tool for representing digital objects and their metadata in a preservation environment. The presentation included the assumption that XML is suitable, more than other tools, for building and maintaining complex logical frameworks through the definition of tags and related grammars and semantics.

The questions listed during the session to which the presentation has tried to provide answers are:

- Options for technical solutions to the preservation of archival records over the long term
- Advantages and disadvantages of XML for archival preservation
- Qualities making XML a particularly suitable method for specific types of digital objects
- Crucial aspects in defining XML schemas for preserving various forms of digital objects
- Usefulness of establishing partnerships in the development and use of XML
- Barriers that will prevent agencies and institutions from using XML

With reference to the abovementioned questions, the presentation underlined the relevance of tagging in order to create and preserve documents, as a syntactical and lexical proposal to solve the increasingly significant interoperability and preservation problems. The criticality and centrality of DTD availability, from the creation phase of the digital resource, was underlined. DTD is able to guide/bind the interpretation of the source, both in terms of the content it aims at conveying and the de-codification tools, and in terms of the arrangement of a reference framework to qualify the message defining the border within which the message can/must be critically analysed. At the same time, the widespread use of DTD and schemas raises some general perplexities on the risk of crystallising documentary forms. This risk is particularly felt at a time of great freedom on the part of the creators in the original creation of the sources. In addition, special attention must be paid to the cost linked to a widespread use of XML for the description of "document behaviours" and the indispensable relations with the critical reconstruction of sources.

The discussion covered many methodological and organisational aspects. In particular, among the various problems raised in the debate and deeply discussed, mention should be made of the XML translation of databases, with reference both to the relational system and the definition of the necessary metadata for preservation purposes. The problems arising in the case of databases whose data derive from sources external to the system and, more in general, when dealing with information borders in case of complex records, have been underlined. Reference was made to the significance of the analysis of document sources in relation to the interpretation of problems linked to their transformation into shared models and also to the cost of this effort vis-à-vis the expected benefits.

The discussion contributed to the definition of a further list of questions and issues that can be answered only through case studies and practical experiences:

1. To what extent cam XML be considered as the only paradigm?
2. What are the (economic) conditions under which it is appropriate to resort to this type of strategies? The existence of shared widespread typologies?
3. Are there different ways to guarantee a limited loss of information and system functions?
4. Can different requirements be identified for public sector and private sector document creators?

5. What metadata contribute to the correct interpretation of data? Are there common areas despite the different needs of each professional community?
6. Can small-size institutions easily adopt the tool?
7. With respect to the different information levels necessary to guarantee preservation (with respect to *data* bit streams are to be preserved, *information* can be maintained when the information tags are preserved, *knowledge* can be preserved if information on procedural, temporal, spatial and structural relations is maintained) can XML preserve the complex information relating to knowledge? Or rather, as shown by the recent experience of the Italian group at the Supercomputer in San Diego discussed in session 3, XML in its present configuration, is not sufficient to describe and preserve knowledge for which different languages and logic formalisms are necessary as demonstrated by the functions envisaged in the OAIS model?

During the discussion it became clear that XML at this stage is a solution widely used on the marketplace because it can face the interoperability challenge providing “easy” agreements on low-profile grammar. The questions relating to a more advanced use of this standard are still open, as proved by the large number of questions raised during the debate.

Session 2: The role of XML as a standard for metadata preservation and metadata exchange

The session focussed on the relevance of XML functions relating to interoperability and document and metadata exchange not only in relation to the digital resources creation stage (Carlo Batini) but also the advanced treatment of already created objects relevant to the preservation of the digital heritage (Enrico Rendina) and also the projects of the National Swiss Archives for the definition, retrieval and management of metadata for persistent preservation (Stephan Heusher). The speakers used their experiences to identify those XML functions that can be used to transpose into a standard digital format those digital objects to be preserved and shared as well as the necessary descriptive information to define their identities, structures and contexts. The initial questions where soon supplemented by other questions on the use of DTD or schemas identifying the structure of digital resources and on the necessary documentation to support and develop the use of XML:

- How should XML structure be defined (DTD/schema) and why?
- What is the minimal documentation needed to correctly interpret an XML file and where should it be located?
- Do we need different standards for metadata preservation and metadata exchange? Why and what are the alternatives?

Many of these questions have been further investigated through other case studies presented in the third session.

In particular, Stephan Heuscher has presented a successful case study with reference to two projects of metadata acquisition conducted by the National Archives of Switzerland. The positive role of XML in metadata exchange at the creation phase has been presented by Carlo Batini, president of the Authority for Information Technology in the Italian Public Administration, who described two recent experiences in this area, related to the legislation and to the records keeping system as defined by Italian law. The project "*Norme in rete"* specifically refers to the effort of defining a general schema to describe the various acts approved by the Italian Parliament, the Government, the Ministries and the Regions. Enrico Rendina has illustrated the complexity and the results of a process aiming at extracting metadata for description of digital objects. All presentations underlined the importance of mark-up languages – and XML in particular "as an open standard, self-explanatory, human-readable, able to be validated automatically" – to define shared data models that can be interpreted over time and without originally sharing the same grammatical and syntactical structures. Mark-up languages can also guarantee data preservation over time and their re-usability with full data integrity.

Many questions - more organisational in nature than technological - linked to XML emerged during the debate:

- Difficulties related to metadata acquisition because of missing or incomplete documentation and generally not standardized and uncontrolled, unknown data formats, complex and/or proprietary and data inconsistency;
- Difficulties related to the definition of system borders in order to ensure metadata acquisition within the system or through different systems;
- Amount of financial resources required;
- Complexity related to the communication process with stakeholders not familiar with data modelling and metadata acquisition;
- Need to define XML schemes for different document formats: word processing documents, spreadsheets, power point, db, etc.

The discussion focussed the general attention and consensus on some main assumptions:

- XML can be considered as a piece of software engine useful for the normalisation and rendering and further finalization of automatically extracted metadata,
- The use of mark-up languages and the consequent transformation of digital objects into persistent objects, can avoid emulation and multiple migration, and more specifically their high costs,
- XML can be viewed as a tool to overcome the centralisation process and develop cooperation strategies by avoiding fragmentation and without loosing consistency,
- DTDs definition approach is becoming more and more relevant specifically with reference to the need for– difficult to achieve – simple and shared document types (basic DTDs, structured DTDs or loose DTDs?),
- XML is a good example of the expressive power of easy tools: the mark-up language paradigm is a low level solution requiring advanced uses to define a specific language able to describe specific objects and contexts,
- XML is also an excellent training tool. It can be used to teach different communities how to develop and design logical data models, in a hierarchical as well as relational way, without necessarily adhering to definite dominions as it occurs with protocol Z3950.

Session 3. XML as a preservation method (critical analysis of case studies)

The very recent experience of the University of Urbino and the Ministry for Finance in San Diego with the aim of transforming the tax income returns forms created in VSAM format into a series of persistent objects able to guarantee quality and feasibility for digital preservation, has already allowed the team involved in the test-bed (and the workshop audience) to verify the positive role of XML. However the experience has also stressed the need to implement the process with other technological and conceptual tools able to describe the complex functions provided by the digital environment (see the presentation by Marco Rendina and Salvatore Costa and the discussion on NARA and SDSC projects by Fynnette Eaton and Reagan Moore).

Another pragmatic and useful experience has been conducted with reference to e-mail messages management, creation and maintenance by the Dutch digital preservation test-bed (see Maureen Potter's report). These reports (further discussed in detail) have allowed the audience to verify the potentials and limits of XML in the concrete activities for digital preservation, mainly when the objects to preserve are complex and rich of contextual (procedural, administrative, technical) information (i.e. the tax returns of the Italian Ministry for Finance). In the case of a large amount of records referred to the same typology, the main problem concerns the need to include automated metadata definition and acquisition within the procedure of record creation (Dutch Test-bed example). The research conducted at the SDSC and at the NARA Centre for Electronic Records tries and implements the basic requirements, including XML formats, to preserve digital records and their metadata and to evaluate further implementation of XML technologies, not yet present on the market. Many questions arose during the papers presentation and the following discussion:

- What capabilities are needed to automate the processes of appraisal, arrangement, description, preservation and access of digital objects (NARA experience)?
- What could be the role of data grid technology and how can information and knowledge be managed (Italian test-bed)?
- What makes XML a particularly suitable strategy for specific types of objects (Dutch digital preservation test-bed)?

The 30-year long experience of Nara's Current Electronic Records System and the difficulties to deal with the digital preservation challenges have been clearly illustrated by Fynnette Eaton (diversity of formats, complexity of objects, high volumes and rapidly changing nature of systems used to create records). The main objective of the recent NARA programmes on the use of XML for the creation of "archival persistent objects" is that of minimising the number of migrations necessary to ensure permanent preservation without introducing new and complex access barriers. The choice was made in favour of mark-up languages and XML because of several general reasons and specific preservation requirements of the National Archives in Washington.

First of all XML:

- Provides non proprietary tools for the digital preservation of different objects,
- It's robust and flexible,
- It' s continuously under development,
- Allows the separate treatment of structure and content,
- Provides rules and procedures for self-description of digital resources.

With reference to the specific needs of NARA, the choice appeared to respond to:

- The difficulty of facing the continuous and uncontrollable development of technologies and languages,

- The need to acquire, select and preserve a large volume of records by using “semi-automatic harvesting” tools,
- The need to develop a “scalable storage” system to deal with the preservation cost and feasibility by means of different tools.

Betting on the possibility of using XML as the lingua franca of the future, NARA set the following objectives

- Develop the tools and components for an archival system of digital preservation,
- Draw up guidelines and standards based on best practices,
- Organise infrastructures, information materials and opportunities to exchange experiences and learn on the topic (workshops, tutorial experiences, etc),
- Promote research paying special attention to the organisation of the information system for digital preservation purposes and the use of advanced knowledge and visualisation management and visualisation tools.

National and international cooperation is at the centre of the project. This approach is particularly suitable for XML because of its interoperability and sharing features intrinsic to the tool and as demonstrated by the first recent experience in San Diego with the participation of the Università di Urbino and the Italian Tax Administration presented to the workshop by Salvatore Costa and Marco Rendina.

The methodological aspects of this project on the construction of an international test-bed have been discussed. The various stages (applied to a statistically significant sample of 50,000 tax returns for the year 1995 using the VSAM (Virtual Storage Access Method) format were:

1. The selection of a significant set to be used during the experiment,
2. The representation of the data model in XML respecting the hierarchical complexity of original relations,
3. The acquisition (ingesting according to the OAIS dictionary) of these data by means of the prototype developed by SDCS, the Storage Resource Broker (SRB), characterised by a relational internal structure,
4. The verification of the consistency and re-use opportunity of the new data with respect to the original information.

A set of conclusions on the usefulness of the tools has been drawn on the first stage of the experiment. In particular, the development of an XML data model

- Makes data preservation possible since digital objects are considered as a bit streams,
- Ensures information preservation as tagged data,
- Does not allow correct and adequate management of knowledge relationship, for example:

    - Data retrieved from other data bases (logical relations with other data sources),
    - Functional relations resulting from complex algorithms (functional /algorithmic data analysis and processing),
    - Procedural data relating to workflow, i.e. data introduced in the system not by the taxpayer but by the administration at different, regular times for assessment purposes, for example.

Many questions emerged from the experiment. They will be analysed both at national level and with the international research community. The first step will be the assessment of the actual need to preserve the “knowledge” accumulated on the record system and the relevant costs and benefits but also the effectiveness of the mark-up languages to achieve the ambitious objective of managing complex relations.

Similar questions were present in the speech by Maureen Potter who described an advanced project on XML treatment of records produced by electronic mail systems during the creation and sending phases. This was a different and very dynamic approach to preservation problems. Preservation was actively dealt with and pursued already at the record creation stage. The model developed by the *Dutch Test-bed preservation project* does not generate perplexities as to XML management of general information and metadata relating to the content and context of records. Yet further investigation is required on the tracking and protection of information relating to the "behaviour" of the records and the relevant procedures as well as the difficulties of introducing complex management requirements to be entrusted with the "good will" of the users. This aspect was not an explicit topic of the workshop but some participants underlined the need for document creation systems to automatically manage tags and metadata on the structure and content of documents – for archival purposes too- to obtain uniformity, quality and efficiency, limiting the active role of individual users and increasing the automatic acquisition process of all records necessary for digital preservation.

The debate brought to the fore the criticality of the software market, and the subsequent panel discussion confirmed it. The software market, although moving to the right direction, provides immature products, both in freeware and commercial programmes sectors. A general agreement was expressed on the need for the community of digital object preservation to encourage IT experts to develop systems with automatic recording functions. This need was highlighted by the complex research activity of the Supercomputer Centre in San Diego that was repeatedly referred to during the debate. The research was then outlined by Reagan Moore during the last session of the workshop because of organizational problems but is closely correlated with the session described above.

Moore presented one of the most advanced and promising projects in this sector focussing on the great potentials of marking up and supporting languages.

At the centre of the analysis are some unavoidable questions often mentioned during the workshop:

- Adequacy of the abstraction levels of these tools,
- Need to design digital archives in addition to defining independent digital objects thereby identifying and managing different levels of information and relevant procedures:
    - Data (*data bit*)
    - Information, including attributes on the origin and context data which make up the source descriptive metadata,
    - Knowledge, including data and information relating to creation processes and source management:
        - Selection (for example data *type*)
        - Acquisition (for example data *model*)
        - Description (for example *compound document, collection attributes*)
        - Preservation (for example physical *file name*)
        - Access (for example *knowledge space concepts*)

For preservation to be successful, effective and feasible by creating –at low cost and in forms independent from technological platforms - *persistent archives,* the implementation of many interconnected mechanisms is necessary, including the evolution of the *storage system* with reference to:

- Migration of media and transformations of application systems
- Access environment (metadata e attributes)
- Presentation problems (*display tool, object data model*)

Digital preservation - Moore recalled- requires the preservation of a large quantity of infra-structural data, which, in the case of emulation, will simply incorporate the operating system able to use the application, while in the case of migration, requires and enables the use of sophisticated tools with new possibilities such as the definition of *data grids*, ensuring:

- Name transparency (*global name*)
- Location transparency (*global identifier*)
- Transparency of access, with authenticity purposes (*authorization control*, *audit trail*, etc.).

The prototype, the already mentioned SRB, is simply a first sep toward self-*instantiating* archives based on the principle of the definition of preservation processes as logical relations and rules embedded in knowledge management systems using XML as reference standard together with advanced *data grid* technologies, with the aim of dealing with dynamic and complex digital objects. The advanced use of XML - already under experimentation- includes the identification and introduction of syntactical (not only grammatical) elements in the tools enabling the annotation of documents. The aim is that of creating process map identifying the transition of status of the objects, translating processes into logical rules but accepting that in order to recognise semantic rules that are often temporary, complex and difficult to classify, different and advanced sets of multi-sector technologies are needed. Yet this type of technology is still immature in normalised forms (for example *domain maps* and process *maps*).

The presentation by Moore was very stimulating and rich in food for thought and triggered a lively debate. During the discussion the many still unexplored challenges in digital preservation have been underlined with the awareness that mark-up languages will certainly play a role in their solutions.

Session 4. XML, digital preservation and the software market: a round table

The market, its immaturity, the lack of attention to the need for long term preservation of digital resources, was often mention during the various sessions and was the subject of a specific panel discussion with software producers.

The round table has been organised with the participation of software houses involved in the development of XML tools for digital object management and preservation:

- The developers of a specific XML tool, "Tamino" developed by Software AG (Massimo Martucci)
- Filenet, a multinational company strongly involved in building EDMS and ERMS (Carlo Stellati, Stefano Gandolfi e Andrea Bongiovanni)
- An Italian company in EDMS/ERMS which pays special attention to XML potentials, 3D Informatica-Extraway (Franco Bazzigotti).

The speakers have been asked to discuss the reactions of the IT market to the use of XML in the past few years and in the future, with specific reference to quality, cost and feasibility of long term digital preservation. The questions asked are very relevant to provide stakeholders with concrete answers and involved the role of XML as the most promising tool for digital preservation, its relevance for application integration and interoperability but also the problems of indexing many fragmented information elements (*schemas*, DTD, data, documents) and the market outlook for XML.

The panel discussion has clearly confirmed the problems existing in the relationship with IT solution providers, something already noticed during the previous debates. IT companies, even the most focussed on quality and development of new products aiming at providing solutions consistent with the use of open standards, XML included, show a very cautious and even passive attitude with respect to the urgent and complex needs of digital preservation and are reluctant to invest on research activities in this sector.

Pending stringent regulations on the protection of the digital cultural heritage - that is not expected to be approved in the binding terms hoped for by the different contributions with specific reference to the public sector- European bodies have a possible and necessary role to play in guiding this difficult transition. In addition the lack of attention and interest on the question of preservation by users and creators of digital resources is clear and worrying. Only a political initiative to increase awareness can change this situation thereby reducing the risk of loosing the enormous quantity of electronic records that is going to be produced by the Information Society.

Session 5. XML and web archiving (European initiatives)

Web archiving is becoming a crucial problem for the development of the Information Society and its memory preservation. XML is only one aspect of the problem. Its complexity is still unexplored and requires the identification of many basic questions. The initiative for the creation of the European web archives within the framework of a European project is an opportunity to discuss some of these aspects. Andreas Rauber has described this opportunity first as a list of unresolved questions. A future ERPANET seminar to be held in Kerkira (May 2003) will be devoted to this issue, whose complexity requires special attention. The list of questions, still unsolved, is long and includes also basic questions, not only those related to the use of markup languages:

- What should be archived? What will be of interest in the future? Do we need everything? All the time? What applications can we address?
- What can we archive with respect to data acquisition: data? Technology? The entire web? Deep web? Meta-information on data, their producers, and the technology used?
- What/how should it be preserved? Original? Appearance? Content? Functions?
- What strategies contribute to the achievement of these goals?
- To what extent can metadata/XML help us to reach these goals?

In particular Andy Rauber underlined the growing importance of web records and the need to explore these resources in order to promote preservation in line with the EWA European initiative. A consortium of 32 Institutions have adhered to the initiative "addressing the whole scope of web archiving including acquisition of information, archives organization and preservation and archives access and exploitation". On preservation the speaker indicated the topics that need further analysis with specific reference to those problems that are intertwined with the advanced use of mark-up languages:

- The need for scalable storage strategies,
- The relevance of metadata preservation as well as the definitions of their nature and structures,
- The criticality of information indexing and retrieval tools by means of the development of advanced descriptive access tools (*information grids*, *semantic web*, *ontology*, *topic detection and tracking*) that are finding their first application and testing in web archiving.

These are all very complex questions and have been discussed in detail. In addition ERPANET is going to devote special attention to these questions in its future training initiatives.

Conclusions and recommendations

Seamus Ross has described the complexity of the new challenges very well in his conclusions: computerisation provides new problems but also new solutions and prospects. Therefore preservation specialists are compelled to reach new more ambitious goals such as that of handling/preserving/describing not only objects but also the behaviours connected to the production, treatment and use of sources. This is a possible yet critical goal.

Several digital preservation questions - to which XML can provide useful solutions- have emerged during the presentations and the debates, although concrete solutions were only partially outlined. Ross in particular listed the risks and the needs, which are paramount in long-term digital preservation programmes:

- Insufficient existing *encoding* tools and products with respect to XML potentials,
- Need for more experimentation, even with uncertain outcome, and more knowledge on the advanced use of the tool as indicated by Reagan Moore,
- Lack of training and awareness-raising initiatives for producers on digital objects management tools, mainly in the case of very complex resources,
- Need for developing knowledge *management* and representation tools, through the development of a shared vocabulary among the different professional communities,
- The need to encourage the use of XML in its most advanced and complete forms- including the description of behaviours and processes – in view of the difficult and expensive retroactive retrieval of new records,
- Need to focus attention and efforts on the analysis of the benefits connected to a correct creation/maintenance of digital resources, from the financial point of view in particular,
- Importance of training expert cultural operators in the sector of preservation to increase awareness about the risks and potentials entailed in digital environment,
- Need to modify the organisation approach of digital records and objects producers to promote rapid and effective initiatives.