erpachat

Summary of an online discussion of metadata for digital preservation.

6 November, 2003

INTRODUCTION

The very first erpaChat took place on November 6, 2003. It is a chance to 'chat' about topic, not to discuss and posit positions within a structured environment such as a conference. This first chat was set up as a post event discussion forum for the Marburg Metadata workshop that ERPANET held in September 2003. It gave participants and speakers the opportunity to discuss issues that there just was not enough time to cover at the event in Marburg. Speakers from the event very kindly gave up nearly two hours of their time in order to answer questions that had been posed by workshop participants. For this first erpaChat, we deliberately kept numbers quite low and in all, nine people logged on to the session. Discussion was lively and led by questions that had been received in advance of the session.

SUMMARY

The session began with a question directed initially at Steve Knight from the National Library of New Zealand. Steve was asked about the Library's progress in the automatic capture of metadata, and how much could be created automatically in the future.[1]

It is universally accepted that the automatic creation of metadata is necessary to bring down the cost of this expensive and time-consuming practice. Steve reported that the work of the National Library of New Zealand suggests that 80-90% of metadata could be automatically created. To date they have been working with five major file types (Word 2, Word 6, TIFF, BMP and WAV). Changes in the Header structure in different versions of Word required development of separate 'adapters' that are slightly different for each version. Further discussion revealed that file types such as XML and HTML may prove problematic so far as they do not contain header fields.

The main focus of the work the Library is carrying out is the work on preservation metadata, but in terms of automation, there is no reason why the tools for extracting preservation metadata could not be used for descriptive metadata. Knight reported that they are also experimenting with bundling different types of metadata together, rather than keeping them separate. This led onto discussion on elements of overlap in metadata, and reference to the RLG that is looking at this with manufacturers of imaging equipment.

The next question was for Heike Neuroth from the State and University Library, Germany, and covered interoperability and what an interoperable layer model may look like. Heike clarified that the layer model she has been working on has three levels: 1. transport of metadata (e.g. Z39.50, LDAP); 2. representation and exchange of metadata (METS, XML); 3a. attribute space (e.g. DCMES); 3b. value space (DDC). In general, the more standardised services are, then the more interoperable they can be. However, it was noted (and with some reservation) that it also possible to export to standards thus negating the need to be standardised. Under this area crosswalks are key, but participants cited a number of worries about crosswalks.

[1] The New Zealand model can be downloaded here: http://www.natlib.govt.nz/files/nlnz_data_model.pdf.

While they are sometimes useful, it is far more sensible to use common elements. Of course, this is more easily achieved in repositories and services that agree on standards than in a more 'open' sphere where this is harder to control.

In terms of preservation metadata there was agreement that more work needs to be done in order to find a suitable set of guidelines: NZ had done some work in this area, but there was a tension between finding a core set of metadata, and the depth required for preservation and other schemas. Work is also being carried out by OCLC/RLG.[2] The need for a core set has at its foundation the problems of domain specific metadata. Different domains – be they libraries, museums or archives – all have different requirements for their metadata sets. This is due to the different types of object types and also the different user requirements within these domains.

This led naturally onto the semantics of schemas. Are the semantics dependant on the object that is being described? There was agreement that they were, but also suggestions that descriptions are usually put in place to help users discover the object. Importantly, users search in different ways dependant upon what the object is. This key point must be borne in mind when developing metadata.

It was a recurring theme that models have to be in place to enable interoperability. These models have many factors to take into consideration; they have to be cross domain, relevant for both global and local level, be general enough for varying semantics of different objects, yet defined enough to actually prove useful and functional. No small feat, and one which should be explored by a cross-sectoral research group.

The next question asked whether the semantic web could perhaps act as a driving force for creating core preservation metadata.[3] One of the key areas here was the need for provenance in the semantic web, the proof that the information is correct and from a trusted source. [4] In this way, and also in terms of repositories and registries, the idea of the semantic web does have parallels in the discussion. However, the conversation on the semantic web was carried out with an air of doubt and hesitation of whether the semantic web will actually come to fruition.

Finally, the participants tackled the distinction between libraries and archives, and whether cooperation and reduction of possible redundancy of effort was possible. There was suggestion that a baseline of technological sharing would perhaps be feasible. Steve Knight argued that at the repository level, certain file types do not need any initial special treatment.[5] Varied points were made on this area; the question of authenticity and the clients' perspective were raised as factors possibly limiting this sharing. The key issue was where exactly the shared layer is in the structures, and where diversity took over. However what was agreed was that both communities

[2] For this work see: http://www.oclc.org/research/projects/pmwg/default.htm.

[3] Participants were pointed to the SIMILE project (http://web.mit.edu/simile/www/) that looks at semantic interoperability of metadata.

[4] Michael Day pointed out that provenance in this context was not used in its archival sense, and meant knowledge of where the metadata had come from, and also included some indication of trust and context.

[5] The example used was pdf.

should consider a common approach in certain base areas, before developing their own separate methods.

The session closed almost two hours after it began, and all participants agreed that the erpaChat had been a success and a fun and novel approach.

A similar event chat is planned for the Rome Workshop.

PARTICIPANTS

Delphine Jensen, European Investment Bank, Luxembourg
Rienk Jonker, Archives inspector, Netherlands
Jane Stevenson, University of Manchester, United Kingdom
Barbara Hoen, Landesarchivdirektion, Germany
Andrew Wilson, National Archives, Australia
Heike Neuroth, State and University Library, Germany
Malcolm Todd, National Archives, United Kingdom
Michael Day, UKOLN, United Kingdom
Steve Knight, National Library, New Zealand
Thomas Severiens, University Oldenburg, Germany
Wendy Duff, University Toronto, United Kingdom (sabbatical)
Niklaus Bütikofer, ERPANET Director, Switzerland
Pete McKinney, ERPANET Coordinator, United Kingdom
Andreas Aschenbrenner, ERPANET Editor, Netherlands