



MIXED Consultation Workshop

September 10, 11, 12 2009

MEMO 02 July 20th 2009

Contents:

1. Workshop program
2. More on MIXED
3. Participants
4. Venue
5. Reimbursement of costs

1. Workshop program

The main objective of the MIXED consultation workshop is to receive input from experts for the compilation of scenarios for the long-term usability of the components created in the MIXED project. Below the program of the consultation workshop is sketched in broad outlines.

Thursday, September 10th

(Start: 12:00 H with a lunch)

1. Introduction of the workshop
2. Background of the experts
3. Outline of meeting, expected results of the consultation workshop

Friday, September 11th

1. Presentation of the MIXED projects and results
2. Elaboration on the scenario-subjects
3. Break-out sessions in "mixed" groups

Saturday, September 12th

1. Presentation and discussion of the break-out sessions
2. Concluding remarks

(End: 13:30 H after lunch)

Topics to be covered during the workshop are:

- * Evaluation of the quality of the results of the MIXED project
- * Sustainability of the results of the MIXED project
- * Calibration of the outcomes of the MIXED project with other initiatives
- * New directions building on the MIXED project

2. More on MIXED (by Dirk Roorda, project manager of MIXED)

(The intention of the section is to provide the invited experts with background information on the MIXED projects and its results)

The intended results of MIXED consist of several parts: (i) at the heart there are the conversion plug-ins that "know" obsolete file formats and convert the data in such formats into XML. This points at the second key result: (ii) a set of XML schemas, brought together under one umbrella (SDFP = Standard Data Formats for Preservation), that represent the way the information in file formats will be encoded in XML. In order to use MIXED, a repository needs a (iii) piece of adaptor software, so that the repository software is able to use MIXED at designated points in the repository workflow. And last but not least, we hope to contribute to (iv) the propagation of MIXED

as an implementation of “smart migration”, and as a means to practical digital preservation for repositories.

Let me expand a little of some of these results as they are taking shape at present.

(i) Conversion plug-ins. During the last months hard work has been done on hard problems: several binary file formats are now within the scope of MIXED conversion plug-ins. They are: Microsoft Access, the pre-2007 versions; DataPerfect (the WordPerfect database companion) and dBase, versions 3, 4 and 5, by the original vendor and in the Clipper and FoxPro implementations. We used a DOS emulator and original software packages to do a lot of experimenting, together with some open source libraries we found on the Internet. We did not copy the code, but used it for inspiration. The result is 100% pure Java code, publicly usable to convert files in any of these formats into SDFP.

(ii) Standard Data Formats for Preservation (SDFP). We are not smarter than the collective user/developer communities and companies that do databases, spreadsheets, statistical data, and so on. Yet these communities have just begun to express concerns about digital preservation into their formats and software. MIXED's policy is to connect to the most promising initiatives, and bring the corresponding file formats under its umbrella.

That means that for spreadsheets we have chosen ODF (= Open Document Format, used by Open Office). We faced a choice between Microsoft's OOXML and ODF. On the one hand, we trusted the standardization process of ODF more than what was happening around OOXML, and the definition ODF format is much less bulky than that of OOXML. The development of the user base of OOXML, however, could grow to become much bigger than that of ODF. If that happens, OOXML will be a more usable format than ODF from an end user's perspective, and hence from a repository's perspective as well. But it remains to be seen how this kind of usability and ease of preservation will be balanced by actual repositories.

For databases there does not seem an XML format that is canonical. When database systems do XML exports, they usually give the user the opportunity to define and/or customise the XML schema of the export. For MIXED we need a single schema that fits all databases. We have found a schema in which you can express the data model of a database and its data in a generic way. The data model defines the tables, with their names, properties, fields, constraints, and relationships. We have chosen to include the declarative properties, but not the action items, such as triggers and forms.

(iii) We just started specifying points in DANS's repository workflow where MIXED could step in. For the moment we will use MIXED by triggering it from a repository user interface, but ideally MIXED will be scheduled to perform tasks automatically when certain conditions are met and certain files are encountered. There is nothing in the MIXED software that prevents that. But MIXED implements a new preservation strategy, and we need to find the proper place for MIXED by experience. In order to facilitate the invocation of MIXED, we plan to write a piece of adaptor software that calls MIXED as a web service, feeding it with input files and collecting result files.

(iv) We have high expectations of the consultation workshop in September. To us it is important to find scenarios where smart migration is really useful. We also seek advice on how to present SDFP to a public of repository managers, archivists, researchers and developers. And we hope that the use of MIXED spreads into two dimensions: more users and more usability. If we attract the attention of power users such as repositories, with budgets to take digital preservation measures, they are candidate plug-in writers for new file formats and new data kinds. Conversely, repositories that use a growing MIXED can delegate an increasing portion of their file format conversions to MIXED. If this works out, MIXED will collect file format conversions like a snowball.

3. Participants

About 20 people will be invited to attend the consultation workshop. We approached a number of organisations and projects that are relevant for the MIXED project and we are very delighted that

we received positive reactions from almost all organisations. The provisional list below contains the names of experts that will participate in the workshop taking into account that some minor changes in the list might occur.

<i>Classification</i>	<i>Name</i>	<i>Project</i>	<i>Organisation</i>	<i>Position</i>
Digital Preservation Tools	Rainer Schmidt	PLANETS	Austrian Institute of Technology	PLANETS Workpackage leader
Digital Preservation Tools	Adam Farquhar	PRONOM/UD FR	British Library	Head of Digital Library Technology
Digital Preservation Tools	Amir Bernstein	SIARD	Swiss Federal Archives	Project leader / SIARD application owner
Digital Preservation Tools	Marc Kemps-Snijders	CLARIN	Max Planck Institute University	Coordinator User Applications
Research infrastructure	Steven Krauwer	CLARIN	Utrecht	Coordinator of CLARIN
Research infrastructure	Vladislav Makarenko	DARIAH / ESCIDOC	Max Planck Institute	Developer ESCIDOC Project
Research infrastructure	Herve L'Hours	CESSDA PPP DARIAH /	UKDA Oxford University	Implementation Manager Assistant Director Oxford Univ. Comp. Services
Research infrastructure	Lou Burnard	TEI	University	Director of Software development
Scientific data archives	Nathan Adams		ICPSR 3TU	
Scientific data archives	Jeroen Rombouts		Datacentrum Tilburg	Project leader
Scientific data archives	Rob Grim		University	Research Data Specialist
Outreach of open source software	Arofan Gregory	DDI	ODAF	Manager ODAF
Outreach of open source software	Merce Crosas	DataVerse	Harvard University	Director of Product Development
	Dirk Roorda	MIXED	DANS	MIXED Project manager
	Jan van Mansum	MIXED	DANS	Developer MIXED project
	René van Horik	MIXED	DANS	Dissemination MIXED project
	Peter Doorn		DANS	Director DANS

4. Venue

The consultation workshop will be held at the Carlton Beach hotel at Scheveningen <<http://www.carlton.nl/beach/default-en.html>>, located on the shores of the North Sea. The hotel can be reached by public transport within 1 hour from Schiphol airport (Amsterdam). The next memo will contain more information on the hotel and its location.

5. Reimbursements of costs

The costs for travel, accommodation and meals are covered by the Mixed project. All participants can claim expenses based on a form that will be available at the consultation workshop. Please let us know if you prefer another procedure.

For more information on the MIXED project, contact Dirk.roorda@dans.knaw.nl (on holiday until August 20th)

For information concerning the consultation workshop, contact Rene.van.horik@dans.knaw.nl, tel. +31-6-23297389 (on holiday until August 12th)