The International Journal of Digital Curation
Issue 2, Volume 6 | 2011

Migration to Intermediate XML for Electronic Data (MIXED): Repository of Durable File Format Conversions

René van Horik and Dirk Roorda,
Data Archiving and Networked Services (DANS – KNAW)[1]

Abstract

Data Archiving and Networked Services (DANS), the Dutch scientific data archive for the social sciences and humanities, is engaged in the Migration to Intermediate XML for Electronic Data (MIXED) project to develop open source software that implements the *smart migration* strategy concerning the long-term archiving of file formats. Smart migration concerns the conversion upon ingest of specific kinds of data formats, such as spreadsheets and databases, to an intermediate XML formatted file. It is assumed that the long-term curation of the XML files is much less problematic than the migration of binary source files and that the intermediate XML file can be converted in an efficient way to file formats that are common in the future. The features of the intermediate XML files are stored in the so-called Standard Data Formats for Preservation (SDFP) specification. This XML schema can be considered an umbrella as it contains existing formal descriptions of file formats developed by others. SDFP also contain schemata developed by DANS, for example, a schema for file-oriented databases. It can be used, for example, for the binary DataPerfect format, that was used on a large scale about twenty years ago, and for which no existing XML schema could be found. The software developed in the MIXED project has been set up as a generic framework, together with a number of plug-ins. It can be considered as a repository of durable file format conversions. This paper contains an overview of the results of the MIXED project.[2]

[1] Royal Netherlands Academy of Arts and Sciences (KNAW): http://www.knaw.nl/smartsite.dws?id=25792&lang=ENG

[2] This paper is based on the paper given by the authors at iPRES 2009; received January 2009, published July 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Reuse of digital data is often hindered by non-transparent file format specifications. These formats are often owned by vendors, are dependent on applications, and exist in various versions. Data in older binary file formats are very much at risk now. Obsolescence of file formats is one of the main factors that threaten the long-term access to digital data. In the course of time datasets in a wide range of formats are deposited at scientific data archives such as Data Archiving and Networked Services (DANS) and the future usability of these datasets is obviously a major concern for the management of a data archive.[3] Next to monitoring the state of art concerning digital preservation and the application of relevant tools and standards DANS also undertakes initiatives to develop services that enhance the durability and reusability of its digital assets. The Migration to Intermediate XML for Electronic Data (MIXED) project, that started in 2007, is an example of such an initiative.[4] The MIXED project is based on the outcomes of Dutch digital preservation testbed-projects (ICTU, 2003a; ICTU, 2003b). The aim of the MIXED project is to develop a sound theoretical framework for the curation of file formats and practical services and tools that support this framework. It is the ambition of the MIXED project that its results are not only relevant for the DANS data archive, but also for other repositories of digital data objects. In order to use the project resources in the most efficient way the MIXED project tries to use existing building blocks if possible and to develop new ones in situations where no solutions are available.

The remaining part of this paper consists of four sections. First the *smart migration* strategy is explained as this strategy contains the philosophical basis upon which the MIXED tools are based. Then attention is given to the XML schema with the name Standard Data Formats for Preservation (SDFP) that serves as the interchange format for a number of file format types. The third section elaborates on the MIXED software tools that carry out the smart migration strategy. The MIXED project delivers open source software that can convert various data formats into the XML data format and vice versa. The last section of this paper contains concluding remarks concerning the relevance of the project outcomes for a broader user-community.

Smart Migration

By the year 2000 three main strategies towards digital preservation had been described (Jones & Beagrie, 2001). These were the technology preservation strategy, the technology emulation strategy, and the digital information migration strategy. Another important building block to realise a digital preservation framework is the Open Archive Information System (OAIS) reference model (Consultative Committee for Space Data Systems, 2002) that establishes a common framework of terms and concepts relevant for the long-term archiving of digital data . Based on the principles mentioned above a wide range of initiatives have been undertaken to improve the long-term durability of digital objects.

[3] DANS: http://www.dans.knaw.nl/en.
[4] The website of the MIXED project can be found at: http://mixed.dans.knaw.nl. The site contains a project white paper (Roorda, 2007), see: http://mixed.dans.knaw.nl/background.

Migration concerns the re-encoding of digital information in new formats before the old format becomes obsolete. The purpose of migration is to preserve the intellectual content of digital objects and to retain the ability to retrieve, display and otherwise use them in the face of constantly changing technology. The migration frequency is a function of the rate of change of context. Since the dependency on the context is manifold, and since application software changes rapidly, it is very difficult to keep track of all the migrations that are needed. Another disadvantage of this "plain migration" strategy is the fact that migrating archived material can require a huge amount of work. In addition, it is risky, because it might introduce conversion errors. It pays, therefore, to find strategies to optimise this migration process.

The plain migration strategy keeps data in the formats belonging to the application with which they were created, and thus many different data files follow many different migration paths. For example, in 2010 it might occur that Word documents created prior to Word 6.0 can no longer be opened by the new version of Word. So all pre-1993 documents suddenly qualify for migration.

In order to avoid this situation, the smart migration strategy converts all data files, upon ingest to the archive, into an intermediate generic format expressed in XML. Upon dissemination the file is converted from this generic format into a current vendor format of choice. This is illustrated by Figure 1. It is likely that the intermediate will also change, but at a much slower rate. The optimisation is that conversions are split into many contemporary (or *synchronic*) conversions and a few time-bridging (or *diachronic*) conversions. This is a much more manageable situation, because the complexity of many different formats can be dealt with contemporaneously, and the complexity of bridging time can be dealt with by means of one well defined format. This is illustrated in Figure 2. Smart migration can be considered as a combination of normalisation and migration.

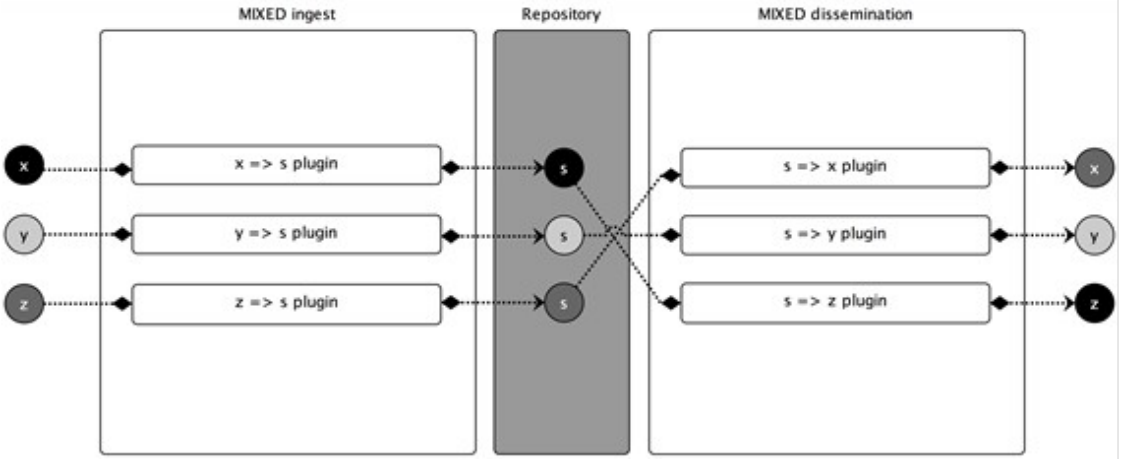

Figure 1. Smart migration. Upon ingest files are converted into an intermediate XML format. Upon dissemination the file is converted to a format of choice.

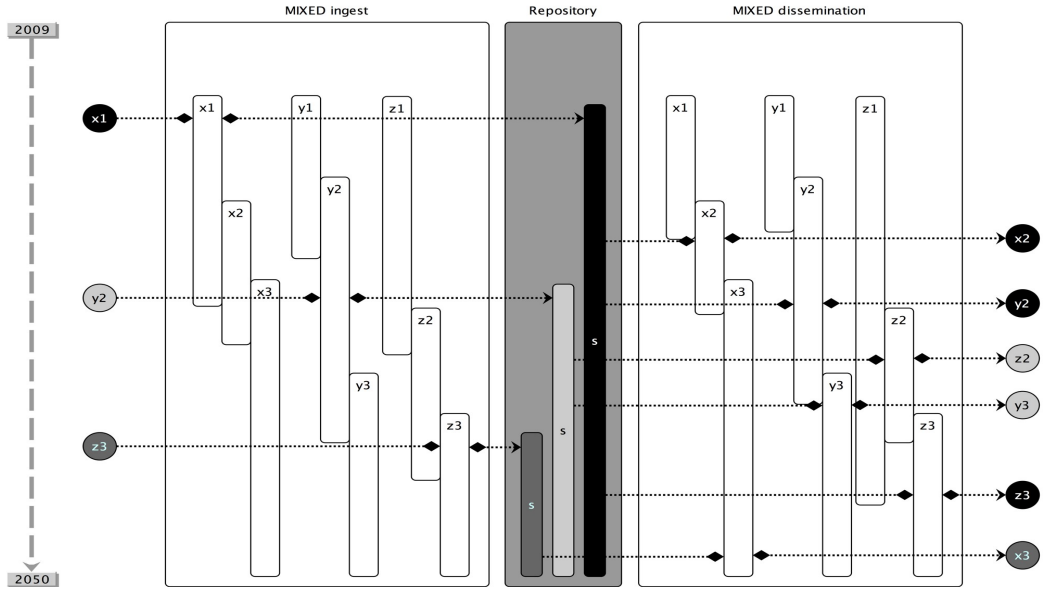

Figure 2. Smart migration. Only a few time-bridging conversions are needed.

A distinction should be made between three manifestations a digital file can have in a digital preservation environment. First, the original copy that will become obsolete for reasons mentioned above, but still serves as an anchor for the authenticity. Next a preservation copy can be distinguished, for example, in pdf format, that retains the original look and feel as a precaution against the loss of aspects that were deemed not worthy of preservation; and last but not least the logical manifestation, typically in XML, that stores the content in a very transparent way, but that may lose aspects having to do with presentation and with action. This is the manifestation that facilitates future reuse of data. Smart migration concerns the processing of files in order to deal with this logical manifestation.

Thus, a smart method of migration is to migrate to data formats expressed in XML, not to a newer version of a non-transparent format. It is then fairly easy to maintain conversions from this XML data format to the various file formats that users want their data delivered in.

The software tools created in the MIXED project demonstrate the viability of the smart migration strategy and are discussed further on in this paper. First attention is paid to the role and implementation of the XML data format in the MIXED project.

SDFP: Standard Data Formats for Preservation

To a large extent the dependence on specific software and hardware for the processing of digital objects can be avoided by using the XML data format. This standard data format is considered self-descriptive and it does not require proprietary software to get access to the data. As XML provides mechanisms to impose constraints on the storage layout and logical structure of the data it is obvious that XML is used to express the generic structures for different kinds of file formats, such as spreadsheets and databases.

Within the MIXED project the XML schema SDFP is developed that defines the features of the intermediate XML data format.[5] SDFP is an umbrella format; it contains sets of XML schemata for various significant data types and builds on existing XML representations of files formats such as the Open Document Format (ODF; see Figure 3).[6]

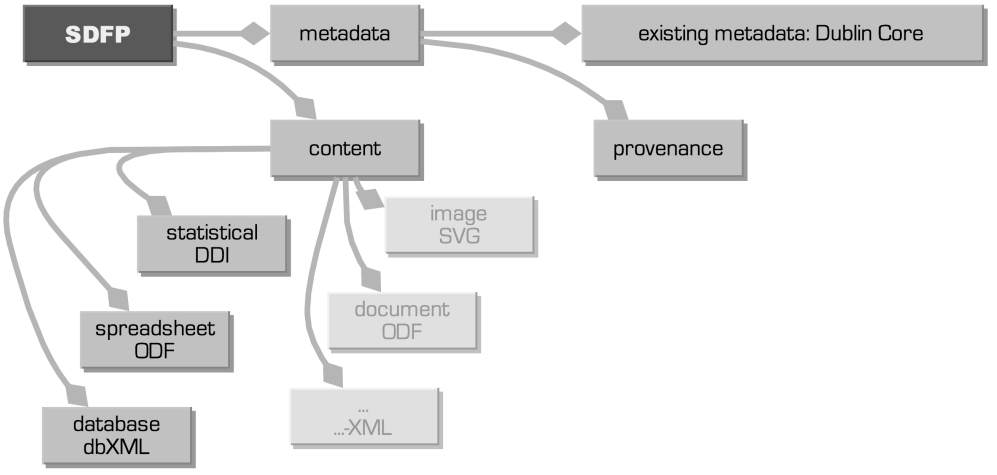

Figure 3. SDFP as umbrella format. DDI: Data Documentation Initiative; SVG: Scalable Vector Graphics; Dublin Core: Dublin Core metadata record.[7]

The SDFP schema is intended to evolve during its use by several organisations. It reflects the aspects of data that are deemed worthy of preservation. SDFP will expand as new data types are added, but it will certainly remain backwards compatible. SDFP can be considered as a device for containing and accumulating knowledge on the structure of file formats and is an essential element in the implementation of the smart migration preservation strategy.

Within the wide range of existing file formats the MIXED project started to concentrate on tabular data in spreadsheets and databases. There are several reasons for this. It turned out that data archives hold an enormous amount of legacy material in spreadsheets and databases that come in many (aged) versions, such as DataPerfect (database), Excel (spreadsheet), Lotus1-2-3 (spreadsheet), Access (database) and dBase (database). As an exercise in extending the scope, we also dealt with data created by means of statistical packages such as SPSS. The importance of these data lie in their content, not in their initial look and feel. This makes these data particularly appropriate for the MIXED framework.

The SDFP representations express the generic structures of databases and spreadsheets. It is the intention that the SDFP schema will gradually be extended with more file types. In addition it can be expected that, over time, the format type descriptions in the SDFP schema will change as it is to be expected that there will be new concepts introduced in the functionality of databases and spreadsheets, which must be expressed in the underlying file format. But we expect that, over time, the SDFP format will change at a much slower rate than the individual binary application formats, which makes the application of the MIXED framework an appropriate option.

[5] In earlier stages of the MIXED project the SDFP schema was called M-XML.
[6] More information on the ODF standard can be found at: http://www.odfalliance.org.
[7] DDI: http://www.ddialliance.org/; Dublin Core Metadata Initiative: http://dublincore.org/.

MIXED Software: Framework and Plug-ins

Having positioned the smart migration strategy and the function and role of the SDFP XML schema to express file formats in the durable XML format we now elaborate on the development of the software tools needed to carry out that strategy. The MIXED software has been set up as a generic framework together with conversion plug-ins.

The framework carries out the conversion workflow and administration. Conversion modules can be plugged into the framework while it is running. Several interfaces can be connected to the framework, including amongst others, a web console, a command line tool, or a web service. The framework itself is published as open source software.

The conversion plug-ins contain the intelligence behind the file format migrations (to and from binary formats and XML representations, based on the SDFP schema). The plug-ins are published as open source software, for maximum usefulness. In addition, new plug-ins for other file formats can be added to the MIXED framework. In this way the conversion plug-ins function as devices that attract the best practices of preserving data that come in various file formats.

The software framework has two focuses: (1) the core functionality of converting binary file formats to SDFP and vice versa, and (2) interfacing with data repositories.

Both focuses remain separable. It is to be expected that the conversion functionality will be useful in contexts other than archiving and preservation, so we should not tie this functionality to a repository context.

On the other hand, the main reason for building the MIXED software is to improve the long-term access and reuse of the objects in the repositories of DANS and those of other organisations with preservation measures.

Implementing the MIXED strategy for a repository poses a number of requirements: The conversions must act on behalf of data producers or on behalf of data managers, which leads to different scenarios and different interfaces. Also, the conversions must maintain provenance metadata, which will be repository-dependent.

Discussion

The MIXED project is a contribution to the community effort of gathering quality tools for digital preservation. This is related to the Preservation and Long-term Access through Networked Services (Planets) project that aims to develop practical services and tools to help ensure long-term access to digital cultural and scientific assets (Farquhar & Hockx-Yu, 2007).

The main idea of the MIXED project is deceptively simple: to convert file formats to the XML data format in order to unwrap the data out of application-bound, closed formats. The unwrapped data are easier to preserve. But then these conversions themselves should be sustainable. A key role is for the resulting XML data files. Support and further development of the SDFP schema is important. The schema refers to other XML constructs that are suitable as standard preservation formats for certain kinds of data. SDFP is intended to be a device for accumulating knowledge on the structure of preserved file formats.

The creation of file format converters is notoriously labourious and difficult. It typically exceeds the resources of individual archives to develop complete conversion packages, therefore there is urgent need for cooperation in this field.

DANS will maintain the web console of the MIXED software for demonstration and experimentation purposes. This web console is accessible via the project website (see Figure 4). The libraries created by the MIXED project team, dealing with a number of file formats will also be made available as open source code.

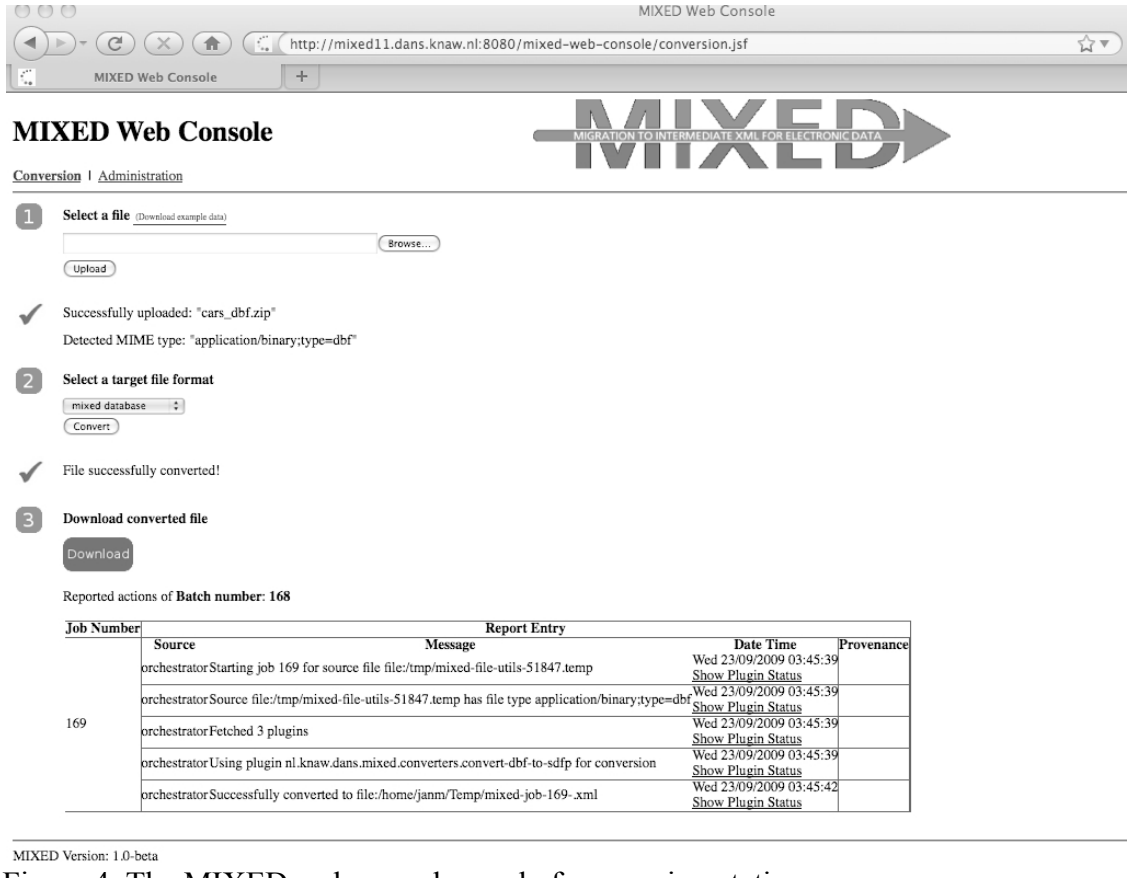

Figure 4. The MIXED web console, ready for experimentation.

DANS will incorporate the smart migration strategy in its data archive management routines. This means that for file formats for which converters are available, XML representation will be created upon ingest and managed using the MIXED software. Upon dissemination, converters will be used to create common useable binary formats. We expect better long-term digital preservation of files in the XML formats that conform to the SDFP schema and requirements. We also expect that the process of converting files to SDFP will come at reduced cost compared to the classic format migration procedures.

The MIXED software will not be able to implement the smart migration strategy on its own. In order to implement an archiving strategy, an organisational process is also needed. Moreover, once the software to carry out the MIXED migrations has been developed, the conversion landscape will need to be monitored for new file formats and other developments. The MIXED software must be able to accommodate a changing conversion landscape. For this, communication and cooperation with other initiatives in the digital preservation community is important.

It is hoped that the methodology and tools developed by the MIXED project contribute to the development and application of a repository of durable file format conversions.

References

Consultative Committee for Space Data Systems (2002). *Reference model for an open archival information system.* ISO 14721. Retrieved February 26, 2011, from http://public.ccsds.org/publications/archive/650x0b1.pdf.

Farquhar, A. & Hocks-Yu, H. (2007, November). Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, *2*(2), 88-99. Retrieved September 15, 2009, from http://www.ijdc.net/index.php/ijdc/article/view/45/31.

ICTU (2003a). *From digital volatility to digital permanence. Preserving databases.* The Hague, Netherlands: Royal Netherlands Academy of Arts and Sciences. Retrieved February 26, 2011, from http://mixed.dans.knaw.nl/qr.

ICTU (2003b). *From digital volatility to digital permanence. Preserving spreadsheets.* The Hague, Netherlands: Royal Netherlands Academy of Arts and Sciences. Retrieved February 26, 2011, from http://mixed.dans.knaw.nl/qr.

Jones, M. & Beagrie, N. (2001). *Preservation management of digital materials: a handbook.* London: The British Library. Retrieved February 26, 2011, from: http://www.dpconline.org/component/docman/doc_download/299-digital-preservation-handbook.

Roorda, D. (2007). *Migration to intermediate XML for electronic data (MIXED). A strategy in digital preservation – A DANS software project* (White paper). The Hague, Netherlands: Royal Netherlands Academy of Arts and Sciences. Retrieved February 26, 2011, from http://mixed.dans.knaw.nl/qr.