Reducing complexity of hybrid data at ingest

Tarvo Kärberg
Project Manager
J. Liivi 4, Tartu
50409, Estonia
+372 738 7585
Tarvo.Karberg@ra.ee

ABSTRACT

The appraisal of records, acquisition and preservation of archival records is regulated by the Archives Act in Estonia. It states that the transferor shall bear the expenses of transfer of archival records to the National Archives of Estonia (NAE), including expenses incurred during re-arrangement, descriptive work and transport of archival records according to the requirements. In practice, these expenses can vary very significantly as agencies and persons performing their public duties have produced archival records in several ways which means that acquisition may require quite a lot of (manual) work. As it may get too complicated and expensive to prepare and transfer data to the archives the producers haven't been too keen to get used to digital preservation.

To overcome the complexity of preparing and transferring the archival records to the archives the NAE has developed a pre-ingest tool – the Universal Archiving Module (UAM). It tries to fix the gap in the current situation by giving producers a set of functionalities which are gathered into one place, are relatively easy to use, can be partially automated and are approved by the archives. Those functionalities include building a classification schema (both manually and automatically from XML files), adding descriptive metadata (both for paper based and digital records, manually or automatically), describing digital content (automatically identifying and characterizing the computer files) and validating (automated control against rules and requirements set by several laws and guidelines) the transfer.

The presentation will introduce some practical examples of how the UAM can help solving complex situations which have occurred in recent practice. The focus is on hybrid data as this is one of the crucial issues on what archives need to tackle. The producers are used to maintain the paper based and digital archival records in a different way, but when it comes to search and access to descriptive information the users would like to get everything from one place. Therefore it is very important to behave proactively at the (pre)-ingest stage and pay very close attention on how the ingest process is being conducted.

Categories and Subject Descriptors

H.3.7 **[Digital Libraries]**

General Terms

Experimentation, Standardization.

Keywords

Digital preservation, knowledge, UAM, ingest, archive.

1. INTRODUCTION

Data providers in Estonia have an obligation to prepare (classify, describe etc.) and transport the archival records to the archive according to the requirements and guidelines set by the archives. Previous experiences have shown that agencies need help to overcome the difficulties that the archiving of records may bring along.

- One of the first obstacles that the agencies encounter is building a classification schema. It may be very time-consuming to perform it manually. They agencies would like to do it in a more automated way.
- The same goes for the descriptions – some descriptions are available in EDRM systems and it would be reasonable to reuse them when describing the archival items.
- Another obstacle may be collecting the content from several sources. The digital content is usually spread between several locations and bringing them into one place for archiving can be difficult.
- One of the obstacles may be finding an easy way to validate your work, check whether you have done everything right when preparing the data for archiving.
- The final obstacle may be the transfer. Agencies want to send their work to the national archives, but they want a smooth, controlled and secure solution for doing that.

To help agencies with those tasks the NAE provides them a special tool – the Universal Archiving Module (UAM) that can help agencies to deal with all previously mentioned obstacles.

2. CLASSIFYING AND DESCRIBING THE DATA

Agencies have often some descriptions available in digital form. Descriptions may be in EDRM or other information systems or even in some Excel files. If an agency would like to reuse these descriptions when preparing their records for archiving then it's should export them somehow from the source system. For example, producing a XML file from Excel is quite straightforward when using Save-As dialogue. The national archive doesn't declare any specific requirements for export. Only one rule should be followed – the export should be in XML format. As the XML format can take various shapes then every system will need a mapping between its metadata elements and UAM input in order to use the UAM import functionality.

Using UAM makes the preparation process smoother as it provides functionality for creating a classification schema and adding descriptive metadata (both for paper based and digital records) manually or automatically. Metadata fields and hierarchy used in UAM are based on ISAD(G), ISAD(CPF) standards and

are therefore commonly suitable for any record's type. More specific metadata elements can be used for lower hierarchy levels (item, computer files). The elements set can be easily extended as each level contains <any> type XML tag.
When descriptions and items contain identifiers then all of them automatically find their right place in the archival classification hierarchy.

3. HARVESTING THE CONTENT

Agencies have often some valuable information encapsulated in computer files and those computer files have not been put into the EDRM systems, but are saved on some network drives.

These computer files are usually organised in some way for better finding purposes, but they are not automatically related to the classification schema used in an agency. Therefore it is important to create somehow the relations between computer files and appropriate records or files when preparing the data for archiving.

This can be done in two ways in UAM. One option is that the archivist selects computer file(s) or some catalogue in the operation system and UAM imports the computer file(s) to the indicated place in UAM and automatically characterizes and migrates them if needed. This is useful when there is a relatively small amount of computer files to import as it requires quite a lot of manual intervention.

The second way is to use XML files as the list of available computer files for import. The XML file can be created from the console of the operation system by printing out the listing of the contents of a directory. This is extremely useful when the computer files are organised by series or functions of the agency on the network drives.

It is important to note that all actions are logged and it is possible to check whether a computer file is the result of a migration or if it is the original file. As no computer files are being deleted during the migration process, it is possible to repeat the migration (into some other file format) later if needed.

4. CONTROLLING THE QUALITY

Archivists can make mistakes during the preparation process or archival descriptions or classification schema can contain some discords. Therefore, it is natural that the work should have some quality control. UAM provides a three-level validation procedure to the archivist. First, a manual input validation which means that when the data is inputted to some metadata field manually it will be automatically validated. Existence of mandatory values is checked during the saving of the views in UAM. The second level is „forced validation", an additional validation (checking the correctness of the classification three etc.) that is done when "Validation" button is pressed. The final validation will be automatically performed right before the transfer.

Validation levels duplicate some rules, but they are mainly complementing each other. For example, some metadata is not mandatory when the archivist starts the describing process and the absence of it will be not treated as a problem until the archival classification schema has been marked as approved by the national archives.

There are also some rules which only indicate a mistake and are treated as warnings. It means that they are not compromising the transfer, but will be highlighted in the delivery agreement later.

5. TRANSFERRING THE DATA

The transfer process is very important for agencies as they give their work to the archives for validation and ingest. They want to perform this step preferably in an automated way.

There are two ways to deliver information packages to the archives using UAM. First, the classification schema, the descriptions and the computer files can be sent to the archives over the secured channel called 'X-Road' (a secured layer between e-services and databases in Estonia). When sending information packages over X-Road the packages will be automatically split into smaller pieces for the most efficient performance purposes. The system checks that no pieces go lost during the transfer and informs the user about the result of the process.

The second option is to save the information packages to some data carrier and bring them to the archives. This is recommended only to those agencies that do not have a connection to the X-Road channel.

6. ACKNOWLEDGMENTS

The author gratefully thanks the Digital Archives at the National Archives of Estonia, especially the Director of the Digital Archives, Mr Lauri Leht.