

# Towards seamless integration of Digital Archives with source systems (Part 1)

Kuldar AAS<sup>1</sup>, Tarvo KÄRBERG<sup>1</sup>

<sup>1</sup>*National Archives of Estonia, J. Liivi 4, Tartu, 50409, Estonia*

**Summary:** Ingest of records and their metadata from electronic records management applications has been a crucial topic for archives in the last decades. Due to the growing amount of records created new ways need to be found to make the process faster, more efficient and at the same time provide sufficient quality and discoverability for the future users.

This paper introduces a solution from the National Archives of Estonia which aims to automate the task of reusing records management metadata for archival purposes. The developed principles and tools rely on an automated metadata mapping solution which allows standardising different metadata sets used in agencies into the central metadata model applied in the digital repository. The paper presents practical experiences and problems of such a model which are additionally discussed in Part 2 of the current paper.

## Introduction

For decades the common practice in archives has been the manual description and arrangement of records delivered by agencies. However, in recent years the use of IT tools, especially electronic records management (ERM) applications, has become widespread and along with that the expectations and needs of users and agencies towards the archives have grown:

- Users are demanding online, single item level access as they are already used to this in the original ERM applications;
- The use of IT has also brought an explosive growth in the amount of records created in archives, therefore also the number of records to be preserved long-term is growing;

This means that archivists in both agencies and national archives need more sustainable solutions, which would simplify the traditional tasks of archival description and arrangement in the case of electronic records and metadata.

Most national archives (as well as the National Archives of Estonia) have concluded that the most reasonable way of dealing with the problem is to look for ways how to reuse automatically original records management metadata to both speed up the archiving processes and raise the quality of archival metadata.

However, looking at the task more closely we can identify multiple issues which need to be solved first. One of the main problems is the poor availability of records management related interoperability standards, incl. metadata frameworks. While a national records management metadata standard has been available in Estonia since 2006, it has not been widely implemented. Furthermore, the central metadata schema is only able to maintain a limited set of central records management metadata elements and each of the agencies implementing it usually amend it with more detailed, agency or function specific, elements. Of course we have to remind ourselves also that even if we achieve the full standardisation of records management metadata now then we have still another 10-20 years when we have to deal with the ingest of records created prior to that. And, last but not least, there are always some private sector companies and private persons who are interesting as data providers for the national archives but do not follow the regulations posed on the public sector.

On the technical side another concern is the multitude of ERM applications available on the Estonian market and implemented in agencies. According to a survey from spring 2011 there are currently 11 different solutions implemented, with none of those having more than 1/3 of market share. Along with the lack of technical and semantic interoperability standards this has created the situation that each of the systems has a separate way to deal with the export of records and metadata, thus in each case a different technical approach of reusing the metadata and importing records to archival repositories is required.

### **Current solution**

In 2006 the National Archives of Estonia carried out a needs analysis which was formulated into the following requirements of a software tool:

- The archives need to provide a stand-alone tool which could be implemented in all agencies and can be used by the agencies' archivists and records managers;
- The tool needs to support a wide variety of different ERM application export structures in an XML format;
- The import functionality of the tool needs to be easily adaptable to changes in the ERM applications;
- The tool needs to allow manual quality checks and updating of records management metadata;
- The core of the tool must be based on the national records management metadata schema but at the same time allow for additional agency and function specific metadata import;
- The tool must automate the process of creating archival descriptions (including single-item descriptions) based on already available metadata;
- The tool must support and automate the identification, characterization and extraction of technical metadata for various file formats;

- The tool must support migration of computer files into preservation formats accepted by the National Archives of Estonia.

In 2008 a software tool called UAM<sup>1</sup> – Universal Archiving Module – was delivered based on the needs. Subsequent major updates were developed based on first test results and the tool was taken into actual use in 2010. The main components of UAM are visible on Figure 1.

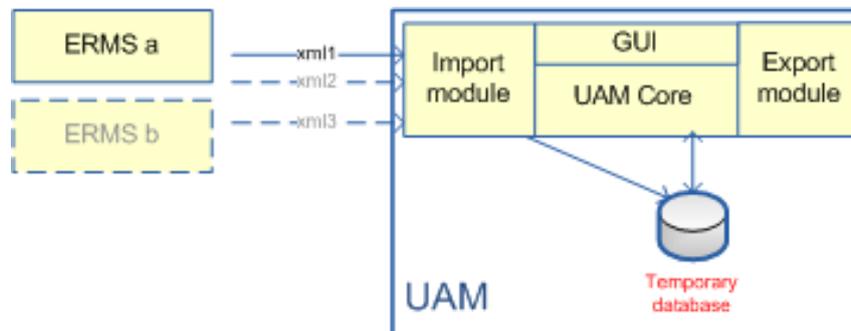


Figure 1: UAM general architecture

### Import module

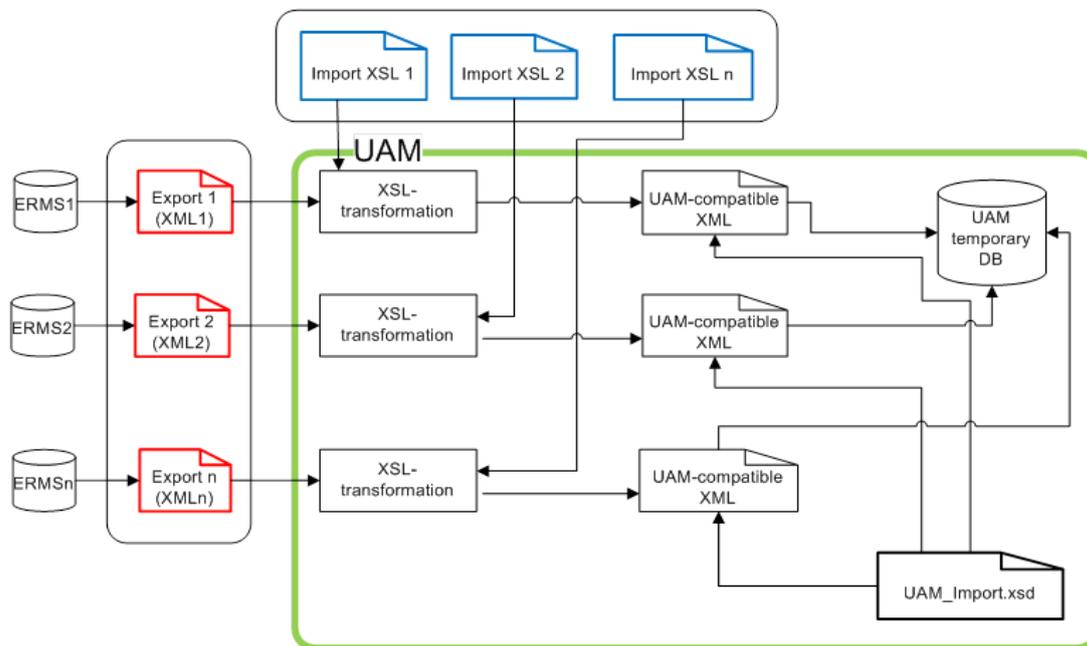
The UAM import component makes use of a native input schema (XML Schema) which is mainly based on the archival description standards ISAD(G) and ISAAR(CPF). Additionally it makes use of some metadata elements originating from the Estonian national records managements metadata set, and also allows additional uncontrolled tags to be included. To be able to cope with XML extracts following different structures and semantics, UAM uses a XSLT engine, which allows administrators to “teach” UAM to recognize the agency-specific extracts. A standardised XSLT transformation is also made available for applications using the standardised national records management metadata schema as the means of exporting records and metadata

During the implementation phase of the UAM at the agencies a two-step configuration should take place:

- the ERM application provider creates a structural mapping between the records management system export format and the native UAM schema. Ideally this action should be done only once for each different records management system;
- the agency’s records managers create a semantic mapping between the metadata elements in their specific records management system’s implementation and the UAM input schema. This action is done separately for each different agency or records management system implementation.

<sup>1</sup> <http://rahvusarhiiv.ra.ee/en/universal-archiving-module/>

This two-step approach should allow limiting the effort needed for the mapping during each new configuration of UAM in the agency and thus maintain a good balance between the required effort and gain in quality.



**Figure 2: UAM import module**

## UAM core and GUI

The UAM core implements all relevant technical and archival requirements, primarily validation rules for archival metadata, and checks the file formats in an easy to use graphical user interface. The user is able to validate the imported metadata for gaps, input missing metadata elements, identify and characterize imported records' components (computer files) and compare them against a list of allowed archival file formats, automatically create technical metadata and if possible, UAM also automatically converts non-conforming files into archival formats. All actions on metadata and file formats will be logged and the users can create reports on the current status of work at any time. If the necessary requirements for archival and technical metadata and file formats are met it is possible to create submission information packages (SIPs) for transferring those to the long-term digital repository of the National Archives of Estonia.

## Export module

The standard installation of UAM creates transfer packages following a XML format defined by the National Archives of Estonia. More explicitly the transfer package consists of two different types of XML files:

- XML file for the archival structure and descriptions of the transferred data , it is expected that each transfer includes one such “table of contents” XML file<sup>2</sup>;
- One XML file for each record including the record level metadata, computer files included in the record and technical metadata about the computer files<sup>3</sup>.

However, as some organizations might want to use UAM for transfers into other long-term repositories a XSLT engine is available to “translate” those native XML files and to export records and their metadata into other transfer formats (i.e. METS<sup>4</sup>).

### **First results**

During the last two years UAM has been actively used in records transfers for both electronic and paper records (i.e. situations where only archival descriptions are created in UAM without accompanying computer files). The feedback from agencies can be summarised as follows:

- Rather surprisingly the agencies see the main gain of UAM and accompanying guidelines in the possibility and need to organise the agencies’ archives in a more detailed manner. This means that using UAM forces the agencies to think through all the different types of records and their metadata as well as the changes which have occurred over the years. Additionally they have to compare all the different description sets to the central records management metadata standard which, all together, allows gaining a better overview of the description status of the ERM application for the records managers;
- The first time implementation and customisation of UAM is still rather time consuming and is in that sense comparable to the duration of the earlier manual description and arrangement processes. However, the agencies think that the tool will have considerable effect during future transfers when the setup and customisation steps do not need to be repeated;
- The learning curve of UAM is said to be low which enables archivists to spend most of the time on customising mappings and preparing transfers instead of learning how to use the tool. This is especially important as such a tool is only used every few years (for each transfer) and not regularly.
- The agencies as well as archivists in the National Archives of Estonia highlight that the possibility to transfer item-level metadata is one of the main gains from using the tools as this will provide the ability to build online tools to search and use single records by both researchers as well as agencies;
- On the negative side, it is apparent that the centralisation efforts in the field of records management metadata are rather strongly opposed by agencies, especially as the

---

<sup>2</sup> XML Schema available at:

[http://rahvusarhiiv.ra.ee/public/Digiarhiiv/UAM/UAM\\_Eksport\\_arhiiviskeem\\_v2.0.xsd](http://rahvusarhiiv.ra.ee/public/Digiarhiiv/UAM/UAM_Eksport_arhiiviskeem_v2.0.xsd)

<sup>3</sup> XML Schema available at: [http://rahvusarhiiv.ra.ee/public/Digiarhiiv/UAM/UAM\\_Eksport\\_arhivaal\\_v2.0.xsd](http://rahvusarhiiv.ra.ee/public/Digiarhiiv/UAM/UAM_Eksport_arhivaal_v2.0.xsd)

<sup>4</sup> <http://www.loc.gov/standards/mets/>

central metadata set is not able to fulfil all their needs and is rather expensive to implement in actual ERM applications controlled by commercial providers. While some hope is set into European efforts like the DLM Forum MoReq<sup>5</sup> initiative it is strongly recommended to look into additional ways of achieving simpler mapping and transfer solutions.

### **Next steps**

To overcome the problems highlighted in the previous chapter the National Archives of Estonia is continuing to look into additional tools and methods to further improve the import possibilities of UAM.

One of the most interesting possibilities currently discussed is the use of semantic annotations based on a records management ontology instead of requiring the agencies to map their metadata elements against central schemas. This approach would potentially help the agencies to overcome their current main problems – as semantic annotation of metadata is happening on top of the actual description activities and is not replacing it then:

- It is potentially cheaper to implement in ERM products;
- It allows agencies more easily to build a metadata schema which allows them to create and maintain the information they need for their daily work;
- It is possible to easily use multiple ontologies (i.e. a records management ontology and a medical or legal ontology) thus allowing to more easily archive record-type specific metadata so that it is also easier to use and connect to external sources during long-term preservation.

However, while the creation of ontologies and implementation of semantic annotation and mapping has been active in the realm of business databases it is not that well advanced in the records management world. Therefore there is a need to first work on the ontology and records management specific annotation practices until this scenario can become a reality.

Of course the idea itself is not new. The Clever Recordkeeping Metadata project<sup>6</sup> has developed a similar minded proof-of-concept metadata broker already in 2006. However, the conclusions of the project highlighted among other things that:

“Current recordkeeping metadata standards lack semantic precision, and canonical machine processable encodings, both of which inhibit their uptake”<sup>7</sup>

Looking at the current developments in records management standardisation, especially at the availability of multiple semantic interoperability initiatives world-wide the question for us remains: would not it be a good time to relive the CRKM metadata broker as an international effort by the world-wide records management community?

---

<sup>5</sup> <http://www.dlmforum.eu/>

<sup>6</sup> <http://infotech.monash.edu/research/groups/rcrg/crkm/>

<sup>7</sup> <http://infotech.monash.edu/research/groups/rcrg/crkm/outcomes.html>

Another interesting approach to solving the problems highlighted in this paper is also presented in the accompanying paper “Towards seamless integration of Digital Archives with source systems (Part 2)”. While the current paper discusses the possibility of solving the mentioned problems mainly during the pre-ingest stage, i.e. outside the archives, Part 2 is concentrating on solutions inside the archives repositories. Our belief is that the future transfer, preservation and access solutions benefit from both of the approaches and thus offer a reasonable level of standardisation supported and amended by a flexible preservation system.