Integration of Records Management and Digital Archiving Systems: What Can We Do Today?

Pauline Sinclair	Alan Gairey	Robert Sharpe
Preservica Ltd	Preservica Ltd	Preservica Ltd
26, The Quadrant	26, The Quadrant	26, The Quadrant
Abingdon Science Park	Abingdon Science Park	Abingdon Science Park
Abingdon, OX14 3YS, UK	Abingdon, OX14 3YS, UK	Abingdon, OX14 3YS, UK
+44 1235 555511	+44 1235 555511	+44 1235 555511
pauline.sinclair@preservica.com	alan.gairey@preservica.com	robert.sharpe@preservica.com

ABSTRACT

Archival organizations can receive digital content from a wide variety of sources. However, the single most common is probably records management systems. The interface between records management and digital archiving systems remains one where there is considerable impedance for a number of reasons, in particular:

- Records need to be appraised. Should this be done within the records management system or after export?

- The files in which records are manifested may become obsolete before the records have reached the end of their active lifetime. This implies some form of digital preservation is needed. Which in turn implies a role for the functionality of digital archiving systems for active records whether or not these records are of long-term archival merit.

- There are a number of technical mismatches between systems. In particular:

    o There is no single export format for records management systems so it is not clear how to export them. This includes how to export logical structures (e.g., hierarchy of records), physical structures (e.g., arrangements of files that are needed for files to work coherently in a given technology) as well as metadata.

    o There is no single metadata schema for records management systems. Instead, structural, descriptive and technical metadata may be described in arbitrarily complex ways.

    o There is no universally agreed import format or metadata schema for digital preservation systems.

- The need for digital archiving systems to receive input from a number of heterogeneous systems.

Standardisation, through projects such as e-Ark, may eventually produce solutions by solving all or a number of the above points. However, such programs will last many years and adoption of any outcomes will take many years more. In the meantime, records will still need to be received and processed by archival organisations. What is the best way of proceeding today? Can a few simple things be done in the short–term to make things easier?

This paper will look at each impedance point in turn and discuss what can be done today and what could be done in a year or two with limited effort.

Categories and Subject Descriptors
D.2.12 Interoperability

General Terms
Design, Experimentation, Theory.

Keywords
Digital preservation, records management, integration.

1. INTRODUCTION
Typically archives have many sources of content, but probably the most common is records management systems. This holds true in both the analogue and digital worlds. In the analogue world, the traditional model for integrating records management and archiving is that archiving is something that happens at the end of a record's life. Archiving is one disposal option, alongside destruction and transfer to another records management system or organisation.

While the traditional, **end-of-life archiving** model has the advantage of being simple and easy to implement, it has the disadvantage that in the digital world the files in which records are manifested may become obsolete before they are archived to a digital preservation system. This is only a problem for records with long retention periods, i.e. "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community" [9].

2. DIGITAL PRESERVATION OF ACTIVE RECORDS
To ensure continued access to digital records with long retention periods, it is necessary to integrate digital preservation with electronic records management, but how? There are three possible models:

- **Synchronised archiving**, where all records are held in parallel in a digital preservation system (DPS), as well as the electronic records management system (ERMS).

- **Syndicated archiving**, where all content (records and archived material) is held in the ERMS and the DPS just holds preservation metadata and initiates and provides preservation services (characterisation, integrity checking, and preservation planning and action).

- **Integrated archiving**, where the DPS acts as the 'back end' for the ERMS, providing record storage in addition to preservation services, while the ERMS handles capture and access.

While all three address the technological obsolescence of records, the complexity of the solution decreases from first to last. In the synchronised archiving model, records management functionality (retention periods and legal hold in particular) has to be added to the DPS, and coordination between the two systems is required on record capture, security, access, legal hold and disposal. In the syndicated archiving model, the ERMS has to be modified to include some digital preservation functionality (i.e. handle multiple copies and representations), and coordinate deletion with the DPS.

Finally, in the integrated archiving model, the records management and digital preservation concerns are separated, thus necessitating fewer modifications to both systems. The ERMS captures records and hands them over to the DPS to store and preserve; when records need to be deleted, the ERMS initiates the process and the DPS implements it. Access to records is through the ERMS, with the DPS merely supplying an appropriate representation; the only direct access to the DPS is for preservation planning and action.

3. TECHNICAL MISMATCHES BETWEEN SYSTEMS

3.1 Export Format

There is no standard export format for ERMSs. Such a format should set out how to export logical structures (hierarchies of records (e.g. Series, File), known as aggregations, that provide context), physical structures (i.e. arrangements of electronic files that are needed for files to work coherently in a given technology), and metadata (providing context and technical information). The physical structure of a record, together with any associated technical metadata, is required to maintain access to the record, while the context is necessary to maintain the understandability of the record.

The International Congress of Archives' principles and functional requirements for an ERMS (ICA-Req) [7], which is the basis for the ISO standard (ISO 16175) for Principles and Functional Requirements for Records in Electronic Office Environments, just recommends using open formats. However, it does provide two examples: the Australasian Digital Recordkeeping Initiative's Digital Record Export Standard [2] and the UN/CEFACT's Record Exchange Standard Business Requirement Specification [11]. Both standards provide a very high-level definition of what a Submission Information Package (SIP) used to transfer records and their associated metadata should contain. However neither specifies the structure of the

records to be exported, and such a specification is necessary to support interoperability. In addition, neither standard mandates the use of a specified metadata schema, which is also necessary to support interoperability.

3.2 Export Metadata Schema

In 2002 The National Archives (TNA) in the UK published a metadata standard for ERMSs [10], with an XML schema to follow the final version of the standard. By the time the schema was issued, TNA had decided to move away from testing and certifying ERMSs, so the schema was never adopted widely.

MoReq2 [4] and now MoReq2010 [5] have defined an XML schema for exporting metadata from a records management system, in order to support interoperability. Unfortunately, neither standard has been adopted widely; although several software vendors have announced their intentions to become compliant with either MoReq2 or MoReq2010 only one has done so (Fabasoft Folio Governance), while another is using the XML schema (Automated Intelligence's Compliant SharePoint for Office365).

As yet, there is no single metadata schema for records management systems, although the MoReq2010 schema could fulfil this role if its use becomes widespread.

3.3 Import Format and Metadata Schema

Just as there is no standard export format and metadata schema for records management systems, there is no universally agreed import format or metadata schema for digital preservation systems. OAIS [9] defines the information that a SIP should contain, but does not constrain its structure. Hence there are as many different SIP implementations as there are digital preservation systems. Examples include BagIt [3] and VERS Standard Electronic Record Format [8], as well as various proprietary SIP formats. Likewise, there are many metadata standards, with PREMIS (preservation), METS (structural), MODS, and Dublin Core (descriptive) being the most common.

3.4 Common Exchange Format

The need for a universally agreed exchange format is clear: not only would it allow interoperability between electronic records management systems, but also smooth transfers to digital archives, which often need to accept deposits from many different records management systems. Such a format would need to specify not only the metadata schema to be used, but also the logical and physical structures of records in sufficient detail for any DPS or ERMS to import any package of records that conforms to the standard, regardless of which software application exported it. In addition, a transfer protocol is required to ensure a smooth and efficient transfer of material.

While the need for an exchange format is clear in the end-of-life archiving model, it is also required for the integrated archiving model in order to ensure that records are preserved correctly. The physical structure of the record must be transferred, together with any associated technical metadata, in order to ensure that technology obsolescence can be addressed correctly, while the context (provided by descriptive metadata and the logical structure of the record) is required to ensure that the record (an information object) is interpreted correctly. The context could be held in the ERMS for as long as the ERMS continues to provide

access to the contents of the ERMS and the DPS, but would need to be transferred to the DPS at the end of the ERMS's life.

The EC-funded E-ARK project [1] aims to develop a suitable exchange format and protocol; however, the project has several years yet to run, so this is not an immediate solution.

4. APPRAISAL

Records need to be appraised for their archival value, but should this be done within the records management system or after export. Traditionally, a records manager would assess which records need to be disposed of to the appropriate archive when setting the retention schedule for a record or aggregation and then on arrival at the archive the records might be appraised a second time to determine whether they met the archive's acquisition policy and so should be accessioned.

Given the volume of electronic records being created in the modern, digital era, pragmatically it is not possible to appraise individual records or archival accessions at the item level. Therefore appraisal has to occur at a higher level, both in the ERMS and on accession to an archive. This has knock-on consequences in that more ephemera, such as personal emails mixed in with business email, is getting into archives, which in turn may affect whether a record series can be opened up (because it may contain sensitive, personal data).

Unfortunately, there is no clear solution to this problem either today or on the horizon.

5. EXAMPLE IMPLEMENTATION

Currently it is only possible to implement the end-of-life archiving model, and an example of this is described in section 5.1. However, this model only ensures the continued accessibility of records once transferred to the archive at the end of their active lifetime; it does not ensure the continued accessibility of long-lived records in the ERMS. Therefore we are working towards implementing the integrated archiving model, and a possibility for doing so is outlined in section 5.2.

5.1 Current Capability

A current example of the end-of-life archiving model is Preservica's ability to ingest document libraries exported automatically from a Microsoft SharePoint server.

Document libraries can be exported from a SharePoint server manually via the Central Administration Console. However, SharePoint also allows retention policies to be set on items in a library. The policy can be set so that, following the expiry of the retention schedule on an item, an action can be triggered automatically. (These retention policy features allow us to consider SharePoint as a candidate Records Management system.)

Preservica supplies a SharePoint solution file that can be installed and deployed via the Central Administration Console. This solution automates the export of a document library, and can be selected as the action to be triggered when a retention schedule expires.

Preservica can be configured to watch the network location to which SharePoint exports document library packages. As soon as a package file appears, Preservica can process it for ingest. The Preservica ingest workflow preserves the folder structure of the document library, as well as all metadata associated with the folders and files within the library.

5.2 Future Capability

A likely candidate solution for the integrated archiving solution is one built around the Content Management Interoperability Services (CMIS) protocol [6].

Many Records Management systems (such as FileNet, HP TRIM, Alfresco, EMC Documentum, SharePoint) already implement this protocol to some extent. The Preservica DPS also implements part of the CMIS protocol. By extending the CMIS implementation in both the ERMS and the DPS, an integrated archiving solution can be achieved, where the DPS can be considered as a remote repository "hosted" within the ERMS.

Content added to the ERMS is then "preserved at creation" by calling the relevant CMIS operations to store the content directly in the DPS.

Several challenges remain for this kind of integrated archiving solution, mainly the fact that the CMIS protocol is built around the synchronous handling of individual files, while a DPS typically deals with the asynchronous processing of SIPs (i.e. related groups of files). Nevertheless, the application of suitable constraints to govern this data model mismatch, plus the ubiquity of CMIS within ERMSs, means that this remains the most suitable candidate solution for integrated archiving.

6. REFERENCES

[1] Aas, K., Bredenberg, K., Delve, J. 2014. Integrating Records Systems with Digital Archives: Current Status and Way Forward. In *Proceedings of International Congress on Archiving* (Girona, Spain, 13-15 October 2014). Retrieved 21 October from http://www.girona.cat/web/ica2014/ponents/textos/id131.pdf.

[2] Australasian Digital Recordkeeping Initiative Digital Record Export Standard. ADRI-2007-01-v1.0. 31 July 2007. Retrieved 16 September, 2014 from http://www.adri.gov.au/content/products/digital-record-export-standard.aspx.

[3] Boyko, A., Kunze, J., Littman, J., Madden, L., and Vargas, B. 2014. The BagIt File Packaging Format (V0.97). Network Working Group. 28 January 2014. Retrieved 21 October 2014 from https://tools.ietf.org/html/draft-kunze-bagit-10.

[4] European Communities. 2008. Model Requirements for the Management of Electronic Records. MoReq2 Specification. ISBN 978-92-79-09772-0. DOI 10.2792/11981. Retrieved 21 October 2014 from http://moreq2.eu/attachments/article/189/MoReq2_typeset_version.pdf.
MoReq2 XML schema retrieved 21 October from http://moreq2.eu/moreq2.

[5] European Commission. 2011. MoReq2010: Modular Requirements for Records Systems. Version 1.1. December 2011. ISBN: 978-92-79-18519-9. DOI: 10.2792/2045. Retrieved 21 October 2014 from http://sysresearch.org/moreq/files/moreq2010_vol1_v1_1_en

.pdf.
MoReq2010 XML schema retrieved 21 October from http://sysresearch.org/moreq/index.php/specification.

[6] OASIS Committee Specification 01. 12 November 2012. Content Management Interoperability Services (CMIS) Version 1.1. Retrieved 21 October 2014 from http://docs.oasis-open.org/cmis/CMIS/v1.1/cs01/CMIS-v1.1-cs01.pdf.

[7] Principles and Functional Requirements in Electronic Office Environments, 2008. Retrieved 16 September, 2014 from http://www.adri.gov.au/content/products/electronic-office-environments.aspx.

[8] Public Record Office Victoria. 2003. VERS Standard Electronic Record Format. PROS 99/007 (Version 2) Specification 3. 31 July 2003. Retrieved 21 October 2014 from http://prov.vic.gov.au/wp-content/uploads/2012/01/VERS_Spec3.pdf.

[9] The Consultative Committee for Space Data Systems (2012). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-M-2. Magenta Book. Retrieved 28 August, 2014 from http://public.ccsds.org/publications/archive/650x0m2.pdf.

[10] The UK National Archives. 2002. Requirements for Electronic Records Management Systems. 2: Metadata Standard. Retrieved 21 October 2014 from http://www.nationalarchives.gov.uk/documents/metadatafinal.pdf.

[11] UN/CEFACT's Record Exchange Standard Business Requirements Specification.23 June 2008. Retrieved 16 September, 2014 from http://www.adri.gov.au/resources/documents/DRES-BRS-20080623.pdf.