INTEGRATING RECORDS SYSTEMS WITH DIGITAL ARCHIVES
CURRENT STATUS AND WAY FORWARD

National Archives of Estonia
Kuldar As
National Archives of Sweden
Karin Bredenberg
University of Portsmouth
Janet Delve

Abstract

While interoperability between active e-government systems has been a significant area of work during the last decade, the fact that much of this information needs to be preserved for the long-term after the initial creation has been ignored. This paper looks into the needs of long-term preservation of digital-born e-government data and describes how the EC-funded E-ARK project proposes further actions to address the challenge of long-term preservation and reuse in a cost-effective manner.

Introduction

The development and application of new and efficient e-government systems has created huge benefits in terms of back-office efficiency as well as in the way that citizens and businesses can interact with government institutions. However, most efforts have concentrated on the active phase of the information lifecycle (i.e. creation and short-term management of data) while little attention has been paid to later stages (long-term preservation and access).

At the same time, national digital repositories (mainly national archives) have the mandate and obligation to ingest, preserve and offer long-term access to valuable pieces of government information irrespective of their format. Unfortunately, necessary tools and methods for handling digital-born information and allowing for similar levels of access users can experience in live e-government systems are lacking at present for long term preservation.

In order to fill this gap, the European Commission has funded a three year project called E-ARK[1] which includes a broad range of leading practitioners from all sides of the issue – records creators and e-government legislators, national archives, research institutes and software providers for both live data and digital preservation solutions.

Current practices and problems

Multiple studies have been carried out in recent years to learn more about the maturity of digital preservation solutions. As an example the Danish[2], Belgian[3] and Swiss National Archives and projects like DC-NET[4], DCH-RP[5] and SCAPE[6] have provided studies on this topic.

---

[1] http://www.e-ark-project.eu/

[2] Kristmar, K. V. (2012): “Common challenges, different strategies.” Retrieved from http://www.sa.dk/media%284227,1033%29/1._Common_Challenges_-_Different_Strategies,_KVK.pdf

[3] Velle, K. (2012): “Database Archiving.” Retrieved from https://www.sa.dk/media(4588,1033)/EBNA-Minutes,_CPH_29-30_May_2012.pdf

[4] Ruusalepp, R. & Dobreva, M. (2012): “Digital Preservation Services: State of the Art Analysis.” Retrieved from www.dc-net.org/getFile.php?id=467

[5] Justrell B., Toller E. (2013): “Standards and interoperability best practice report.” Retrieved from http://www.dch-rp.eu/getFile.php?id=165

[6] Faria L, Duretec K., Kulmukhametov A., Moldrup-Dalum P., Medjkoune L., Pop R., Barton S., Akbik A. (2014): ”SCAPE survey on preservation monitoring.” Retrieved rom http://www.scape-project.eu/wp-content/uploads/2014/05/SCAPE_D12.2_KEEPS_V1.0.pdf

While all of these studies concentrate on some specific issues in digital preservation we can see that in the archival sector practices for preserving digital information are just emerging, and there are only a few countries where digital preservation has indeed been applied in a practical and holistic way. In Europe most notably the UK, Danish, Norwegian, Swiss and Dutch national archives have established relevant procedures and systems to allow for the transfer, preservation and access of born-digital government data. Elsewhere also the US and Australian government and state sectors have been active and successful, meaning that digital-born data is indeed being transferred to archives and can be reused. For other countries and archives the situation is usually less advanced, mainly because of the lack in funding and skilled personnel.

Based on the above mentioned studies, the E-ARK project has continued this work and carried out a comprehensive study in early 2014 to learn more about the technical details of the available solutions[7]. The results of the study show that even in the case where solutions are available, these are rather pragmatic approaches towards the generic problem in that they are limited to only addressing the immediate, pressing necessities of preservation and do not extend to re-use and access. In essence, typical national preservation requirements tend to consist of a set of metadata requirements which must be fulfilled by the agency transferring data, together with some rules for the formatting and structure of the metadata as well as the actual data (i.e. regulations on archival file formats). The normal method of accessing preserved data is through archival catalogues, where users first face the burden of identifying relevant datasets before they are able to start looking for the bits of information they actually need.

One of the main reasons for applying such fragmented approaches is the lack of standardization and therefore also the absence of universal tools which could be applied across borders. Most current solutions are custom built to function explicitly in the legal and organisational framework of a single country or even institution. This means that it is nearly impossible for archives to take up tools developed by others and therefore lower the cost of IT investment. Therefore we can say that the current situation in transferring and reusing digital-born data is as follows:

- each jurisdiction provides its own national standards for pre-ingest and ingest workflows as well as for the Submission Information Package (SIP) structure and content;
- in order to apply these standards, information systems' export functionalities involve custom development by all government institutions, thus making it a significant financial burden;
- the quality of data and metadata harvested from source systems is often lacking due to the limited amount of resources the national archives have for developing relevant quality criteria, according tools and offering training;
- due to the low quality of data and metadata stored in digital repositories it is hard for users to find the pieces of information they actually need access to;
- in most cases the delivery of data to long-term repositories also means that the information is no longer available where the general public has been used to get access to it: inside central mash-up services in national and international service portals running on national interoperability frameworks;
- as such the transfer of data to long-term storage means quite a financial burden as well as a huge loss in accessibility, which makes the data owners less willing to undertake the actual transfer and more inclined to develop their own digital repositories, in turn spending a considerable amount of money to set up systems which do not constitute their core business and without possessing reasonable in-house knowledge of digital preservation.

Need for standardization

From the discussion above it is clear that there is a need for standardizing key elements of the later phases of the information lifecycle. Special attention should be paid to the interoperability steps – actions during which data and metadata is transferred from one system to another, or

[7] The results of the study are available in E-ARK deliverables D3.1, D4.1 and D5.1 here: http://www.eark-project.com/resources/project-deliverables

accessed between these systems. Below we look more closely onto the approach taken by the E-ARK project.

Export and transfer of born-digital data

In terms of standardizing the export of data from source systems, there has already been some effort put into the creation of the MoReq2010 (Model Requirements for Records Systems) specification.[8] The MoReq2010 specification includes, among other parts, high-level requirements for the bulk export of records from systems for records archiving or system migration scenarios. However, the current requirements are mainly sufficient for organisational and partly for semantic interoperability. The level of technical interoperability which would actually allow for the development of universal software components has not been addressed until now.

The goal of the E-ARK project is therefore to build on the high-level MoReq2010 specification and update it by adding more detailed and technical requirements derived from already available national best practices. In particular, the following elements are of key importance:

- a SIP structure that mandates the use of core elements for automating the export, validation and transfer workflows. We can see that the core consists mainly of key elements of technical, structural and administrative metadata which are possible to be created and validated automatically;
- in terms of descriptive metadata the SIP structure must allow the inclusion of any country or domain-specific metadata (as an example metadata specific to eHealth or eInvoice records) metadata to the central core. However, it is intended to create a set of application profiles (i.e. specific SIP variants) for most common types of descriptive metadata (like EAD or MoReq2010 metadata);
- guidelines which describe how the mandatory metadata elements could be created and how to re-use metadata created in e-government systems inside the schema for archiving purposes;
- a pre-ingest and ingest workflow model that outlines the crucial actions of metadata creation; data integrity and authenticity validation;
- transfer mechanisms that allow the bulk transfer of records and their metadata from agencies to archives in an efficient and secure manner.

First drafts of these principles will be available for public consultation as early as January 2015.

As already mentioned the approach is to reuse as many available practices as possible. In terms of export and transfer of digital data we can see that a lot of useful standardisation has already been done outside the archival community. Especially the EC-funded e-SENS (Electronic Simple European Networked Services) project[9] has been developing principles and specifications for data exchange, authentication and the provision of cross-border e-services for the whole European public sector. Obviously E-ARK aims to collaborate with e-SENS to make the outcomes of both projects as similar as possible to lower the cost of development even further.

The outcome of the actions in regard to export and transfer of digital data would in particular allow e-government projects and international software providers (like Oracle, Microsoft and others) to create native data export functionality that can easily be implemented across systems and jurisdictions. All of this will be possible at a fraction of the cost currently put into institutional or national custom developments.

The availability of common specifications will also allow the development of common training and dissemination programs, thus contributing to increased awareness in general as well as to the increase of data quality and understandability in archiving processes.

[8] http://moreq2010.eu/

[9] http://www.esens.eu/

Open preservation formats

Another important aspect to be examined is the standardization and further development of Archival Information Package (AIP) structures and principles. During the first six months of the E-ARK project, a detailed analysis of current prevailing AIP principles has been carried out[10]. This analysis shows that there are already a good number of standards and specifications available as a starting point. For example, PREMIS[11] is widely used for preservation metadata, EAD[12] for archival descriptions, the MoReq metadata module for records management descriptions and finally METS[13] and BagIt[14] for bringing all the different components together. Such an approach is especially visible in the recent AIP specifications provided by the Swedish National Archives[15] and the North-Rhine Westphalia state government[16].

The E-ARK project will continue to evaluate the already available standards and will define a limited core set of mandatory elements for all AIP packages. Most essentially, the elements which are needed for preservation planning, ensuring integrity and authenticity of archived data must be defined to allow for interoperability between different preservation systems.

In addition we aim to add support for additional access-oriented layers to the AIP specification:

- AIP Level 0: for structured data the Level 0 format will allow storage "as is" – with the original data model intact – while allowing for additional semantic enrichment of the contents as an OWL-oriented representation;
- AIP Level 1: the Level 1 AIP is created by analysing the Level 0 AIP and turning it into more easily usable OLAP (OnLine Analytical Processing[17]) cubes, following methods from data warehousing. As such, using a Level 1 AIP will allow the archives to offer easier access to data without the need to learn the specifics of the original data model;
- AIP Level 2: the Level 2 AIP is mainly intended to be used for archiving systems which hold unstructured records (as an example pdf files with common metadata).

These enhancements to currently available AIP formats as well as the tools which will be developed to support the formats, SIP to AIP conversion, and conversion between different AIP levels, will essentially allow archives to store in parallel the original database from which records originate as well as more user-friendly representations (as Level 1 or Level 2 representations) in a harmonised way.

In addition, the possibilities for semantic enrichment of content should be of interest for archives that preserve structured records. Namely, the use of semantic technologies will allow users to search for relevant data by using semantic entities instead of searching for relevant databases and useful elements in their data model. In other words, archives will be enabled to allow searching across database contents independent of their original data models.

Access to archived data

Currently most archives provide access to their digital holdings through dedicated archival catalogues. In addition, the archives tend to organize their content according to archival hierarchical classification schemes and description rules. This means that the "rich" metadata descriptions are usually available only for aggregations of data (collections) but not the single elements (i.e. records) which are mostly the scope of public interest.

[10] The analysis is available as part of the E-ARK deliverable D4.1 at http://www.eark-project.com/resources/project-deliverables

[11] http://www.loc.gov/standards/premis/

[12] http://www.loc.gov/ead/

[13] http://www.loc.gov/standards/mets/

[14] https://wiki.ucop.edu/display/Curation/BagIt

[15] http://riksarkivet.se/publicerade-rapporter-fran-eard (in Swedish)

[16] http://www.danrw.de/?lang=en

[17] http://searchdatamanagement.techtarget.com/definition/OLAP

To give an example, in most countries governmental information systems are currently being archived as database snapshots – the full content of a relational database created in a specified timeframe. This snapshot is usually migrated into open formats and the data model is technically described. At the same time, a content description is usually available only for the whole dataset. As a result, potential users interested in the information must first locate the relevant dataset(s) in the archives catalogue and then query all of these one by one. Added to that, changes in the functions of public sector agencies are also reflected in the scope of their information systems – in the long term the data on a specific topic might have moved between agencies and systems – thus making even the discovery of all relevant snapshots a difficult task.

The E-ARK project is working on a series of solutions in order to overcome these issues. The first approach involves the use of semantic description and data warehousing techniques. The key idea is that if all archived databases were to follow a single formatting specification, it would become possible to apply semantic description and data de-normalisation methods taken from data warehousing approaches (Inmon, 2005). As a result the entry point to preserved data would be simple semantically enriched OLAP cubes instead of relational database snapshots with highly complex data structures. This would allow users to browse the preserved data more easily as well as open new possibilities for data mining on top of the data preserved in the archives.

The other access method which is being researched inside E-ARK is the access to archived records from external systems (as an example government service portals or agency web sites). Again, when we can assume that all government records have been described in digital repositories by using common core metadata elements, it is fairly straightforward to produce API specifications for querying and accessing these records. In more detail, the project is looking at the OASIS standard CMIS (Content Management Interoperability Services)[18]. While CMIS describes the full range of CRUD services (Create-Read-Update-Delete), the application in a digital repository must limit the set to only Read and partial Update services. In addition, the workflow to negotiate for permanent ID (PID) creation and exchange must be examined as well as how to deal with active preservation methods where the technical characteristics of data can change over time (as an example, file format migration might have been applied).

Ultimately, the implementation of the "CMIS Application Profile for Archives" would allow institutions to transfer their data to digital repositories while still being able to continue offering the kind of data access services convenient for their users.

Need for harmonization of knowledge

Despite the lack of standardization, information management (IM) has known extensive research and practice in the past years. In fact, nowadays many business and technical references have emerged to guide the processes of ingesting, managing, preserving and accessing information. In terms of designing the processes, standards such as ISO15489[19] ("Records Management") or ISO30300/1[20] ("Management systems for records") already exist. For implementing tools and services references such as MoReq2010 and ISO16175[21] ("Principles and functional requirements for records in electronic office environments") are well known. For assessing organizations and tools one can refer to ISO16363[22] ("Audit and certification of trustworthy digital repositories") or ISO18128[23] ("Risk assessment for records processes and systems"). Also it is important to note that the examples above are international references. As noted in Section 3 above, due to lack of standardization, several countries have

[18] http://docs.oasis-open.org/cmis/CMIS/v1.0/cmis-spec-v1.0.html

[19] http://www.iso.org/iso/catalogue_detail?csnumber=31908

[20] http://www.iso.org/iso/catalogue_detail?csnumber=53732

[21] http://www.iso.org/iso/catalogue_detail.htm?csnumber=55790

[22] http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[23] http://www.iso.org/iso/catalogue_detail.htm?csnumber=61521

also been defining national practices and procedures that should not be discarded and constitute relevant knowledge to the field.

Apart from that, it is also important that, as a strategy, we admit we live in a world were problems can be seen from different perspectives and consequently solutions need to consider requirements from different areas. Applying this to IM, it means we must recognize other specific views besides only those of information science, such as also information systems, software engineering, risk management, among others. In fact, all this exists already, and is a concern usually known in the engineering and management areas as “Enterprise Architecture”.

The proliferation of standards and references together with the recognition that problems should be analysed from different perspectives, motivations and communities has raised the need for a knowledge system that allows stakeholders to obtain a consolidated view of the existing knowledge. Therefore, one of the goals of the E-ARK project is to design and implement an online open-access knowledge centre that offers the possibility of uploading, managing and consolidating existing best practices, standards and other references not only in the core domain of IM but also in relevant peripheral domains. This service can then be used by business stakeholders in order to understand IM practices and requirements, IM stakeholders as a main source for information and knowledge, and/or, academics and students as a teaching and learning resource.

Valorization of archival data

As a final innovative aspect of the project, E-ARK will promote the re-use of archival data by facilitating a common pan-European approach to providing simple and advanced queries to researchers, to the public and private sectors, and to citizens. The project will research data mining techniques that will enable new forms of data re-use that provide vital decision support for business end users. This will in turn enhance competitive intelligence in the EC digital economy by providing analytical processing access to various longitudinal data sets: demographic, economic, judicial and so on. For example, data mining techniques could be used to compare house price fluctuations across various European cities over time to produce a vital pan-European economic dataset. This could be used as a basis for various marketing and research purposes.

At the other end of the spectrum, archival historical and cultural data could be marketed as part of commercial teaching packages. Data mining within E-ARK will allow analysis of aggregated sets of archival data, comprising structured and unstructured information, to identify new patterns of activity in consumer, business and systems behaviour. It should be possible to analyse open archive records to find trends, correlations, etc. It should be possible to apply post-ingest algorithms to archival records (e.g. automated classification based on machine learning). This ability to analyse activity, rather than survey, sample or observe, is transformational, in that genuine patterns of behaviour can be identified – thus providing a basis for new products and services. E-ARK will thus facilitate the taking up of opportunities afforded by “Big Data” that will become available as a result of archival interoperability.

Summary and timeline

When considering current e-government solutions, we observe that in too many cases the long-term preservation approaches applied form a costly bottleneck in the holistic view of the data lifecycle management.

The E-ARK project plans to change this by providing a set of standardized specifications which can be implemented across borders and therefore allow all archival and other government institutions to apply, discuss and develop these further in a common and collaborative manner. E-ARK should also improve access to archived public information through standard query interfaces and data mining techniques.

While the E-ARK project is still in its early stages, the final results will be available and implemented in a reference implementation by early 2017 and first drafts of crucial elements (e.g. best practice reviews across the archival sector) will be available as early as autumn 2014.

Acknowledgments

E-ARK is an EC-funded pilot action project in the Competitiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

Bibliography

Inmon, W.H. (2005). *Building the Data Warehouse, Fourth Edition*. New York: John Wiley and Sons.