Archives and digital repositories in an eGovernment context: When the subsequent bird catches the worm

Hannes KULOVITS[1], Andreas RAUBER[2], Christoph BECKER[2], Rui GAMITO[3], José BARATEIRO[3], José BORBINHA[4], Milord MAZIVE[5], Domingos JOÃO[6]
[1]*Secure Business Austria, Favoritenstr. 16, Vienna, 1040, Austria*
Email: hkulovits@sba-research.org
[2]*Vienna University of Technology, Favoritenstr. 9-11, Vienna, 1040, Austria*
Email:rauber@ifs.tuwien.ac.at
[3]*National Laboratory for Civil Engineering, Av. Do Brasil 101, Lisbon, 1700-066, Portugal*
Email: {rgamito, jbarateiro}@lnec.pt
[4]*INESC-ID, Rua Alves Redol 9, Lisbon, 1000-029, Portugal*
Email: jlb@ist.utl.pt
[5]*Mozambique's Engineering Laboratory, 1081 Av Moçambique Km 1.5, Maputo, Mozambique*
Email: milordmazive@gmail.com
[6]*Angola's Engineering Laboratory, Rua do Laboratorio de Engenharia, Luanda, Angola*
Email: mangilas28@msn.com

Abstract: Records constitute an important corner stone of governance. As more governments are introducing e-Government solutions, digital preservation turns into an important challenge. This paper reviews both the challenges faced by governments in the area of digital preservation, and scopes several repositories currently installed in different African countries in that context. It then outlines ways to elevate such repositories to long-term digital archiving solutions fulfilling international standards. It points out advantages that making this step now has, potentially avoiding many of the pitfalls that early adopters in Europe and elsewhere are facing, and shows how open-source solutions can be utilized to achieve this goal.

Keywords: e-Government, Digital Preservation, Digital Repositories

1. Introduction

Often leading a wallflower existence to most people, archives undoubtedly play an important role in every society. In particular, state archives all over the world hold important evidence ranging from the state's existence to its very identity. An important example from recent history is the border conflict between Botswana and Namibia which had been settled due to still existent archival maps from the Botswana National Archive [17]. Archives serve an important purpose which goes far beyond providing researchers with the right material for their work at a certain point in time. In his novel Nineteen Eighty-Four George Orwell recognizes the considerable importance of archives [1]: "*Who controls the past, controls the future: who controls the present, controls the past ... The mutability of the past is the central tenet of Ingsoc. Past events, it is argued, have no objective existence, but survive only in written records and in human memories. The past is whatever the records and the memories agree upon.*" Archivists in governmental archives have the difficult task to balance political influence on the one hand, and choose from

records offered to them what is important, interpret, and preserve them into the future. Jimerson attributes significant power over future generations to archives and archivists [3] [4]. Especially for governmental archives, to serve the public's interest in accountability, an archivist's goal is to maintain reliable, accurate and authentic records. With their work they shall represent *accountability*, *open government*, *diversity*, and *social justice* [3].

One of the greatest challenges in conjunction with digital data is that they can be altered easily, without leaving any trace or a clue what the object was like before. Compared to the analogue world a text written on a piece of paper and altered ex post, the original will in some cases be restorable or at least the change can be identified (that it took place) and thus becomes a part of the object. A user who wants to access a digital object at some point in the future must be assured that there have not been any unauthorised changes, neither accidental nor intentional by a malicious manipulation. This means that the digital object must be exactly what it purports to be, i.e. must be authentic, and assert confidence in the identity and integrity of it.

To efficiently manage digital objects and cope with the complexity of multiple producers and consumers of digital information, the concept of digital repository can be adopted. In fact, digital repositories are information systems that store, preserve, disseminate and provide access to its consumers. They are equipped with search and access capabilities not only for information regarding the documents (e.g., title, author), but also for the information contained in the document itself.

With the first recognition of an information economy in 1997, including the government sector [9], much has changed in this field. The electronic services offered by government to citizens, organizations, or for inter-agency communications include for instance, Motor Vehicles Online Registration (Tanzania) [10], and Kenya National Library Service [11]. Repository instances have been deployed in Cape Verde and Mozambique by LNEC, the National Laboratory for Civil Engineering in Portugal. Public national laboratories are another example where digital information is a critical asset for today's governments. LNEC works in the various domains of civil engineering, giving it a unique multidisciplinary perspective in this field. LNEC cooperates closely with several Portuguese-Speaking African Countries (PALOP). Other PALOP countries, such as Angola, are also in the process of deploying a repository instance.

The remainder of this work is structured as follows. Section 2 presents representative initiatives in digital archiving. In Section 3 we argue that digital preservation is a critical activity to assure the continuous maintenance of digital assets, while in Section 4 we give recommendations for setting up and deploying digital archive solutions. Finally, we conclude in Section 5.

2. Digital Archiving in Governmental Institutions

Governments worldwide aim to facilitate interactions with their agencies, citizens and organizations by deployment of ICT – different interactions shall be made easier and more efficient. While these ends can be accomplished through the potentials ICT offers, especially in an e-Government setting trust and auditability of governmental records must be considered in every strategy. Mbwete [10] lists e-Government services which are being implemented in the course of Tanzania's e-Government strategy. In these applications, digital records can be of any data type such as simple text files, audio and video files, image files, database records, or compound document files, each with its unique preservation requirements. Making digital preservation an integral part of any such system will significantly ease the effort necessary to later-on maintain the resulting data accessible and usable. In such systems particular emphasis must be on how the digital records are stored/archived. For archives, information and communication technology offers great opportunities due to the ability of making information easier accessible to a wider audience.

Incorporation of ways to access archival records into already existing e-Government strategies further improves often cited and well-known benefits of e-Government such as increased accountability and transparency [6][7]. However, the challenges of preserving digital material, especially born-digital, are the other side of the coin. While paper can be preserved for many years by sole physical protection, digital objects need constant action on the object itself and the environment required for rendering. Thus, the value of eGovernment services provided by archives highly depends on whether the organisation is able to authentically preserve the digital records and provide means of access to its designated community.

2.1. The LNEC Digital Repository

Civil engineering activities require and produce a heterogeneous set of digital information including, among others, technical reports, scientific datasets, media (e.g., pictures and videos captured in visual inspections), CAD documents representing structural drawings. On the other hand, several users including authority and structure owners are involved in both production and access of this digital information. This information is a critical asset for structural safety, where past information is the best knowledge to predict the future and avoid catastrophic failures including loss of life and environmental damage.

In order to store and preserve critical digital assets, and allow access to the intellectual output of the research and analysis of structural safety control, LNEC added an action line on digital repositories to its planned research programme, which has resulted in the adaptation and deployment of a DSpace[1] based repository.

DSpace, an open-source digital repository platform, was developed in partnership by MIT Libraries and Hewlett-Packard Labs, and serves a variety of digital archiving needs. Currently, a large number of institutions around the world use DSpace to meet their needs, including institutional repositories (IR), learning object repositories, eTheses and electronic records management.

Given the particular civil engineering scenario at LNEC, several adaptations were implemented on top of the custom DSpace platform, such as: new ingest workflows, extension of metadata schemas to accommodate specific required parameters, security options, data validation, frontend customization and multiple import mechanisms. The latest was necessary to contemplate the multitude of local databases and storage spread across distinct organizational units.

Since a considerable amount of documents in the DSpace repository at LNEC are highly confidential, special security and access policies are in place. The platform is only accessible from within the intranet. However, this repository was not conceived with strong digital preservation requirements, lacking preservation plans and adequate actions to fulfil the future understand-ability of the underlying digital assets. On the other hand, and to accommodate the need for scientific knowledge dissemination, an indirect connection with the outside is being established through the Common Repository project (RCAAP[2]).

2.2. PALOP Repositories

The issues of managing digital assets are not exclusive to the civil engineering domain, but a generic concern instead. Given the use case success of LNEC's digital repository, and the strong collaboration with PALOP countries' homologous institutions, as well as their similar

[1] http://www.dspace.org/
[2] http://comum.rcaap.pt/

requirements on information management, LNEC is currently providing the means and support in the installation and deployment of DSpace repositories at the Civil Engineering Laboratories of Cape Verde and Mozambique. Although the requirements are similar and the types of digital assets are the same, some particular concerns had to be taken into account as, for example, document access policies and security rules. For instance, while LNEC's instance is inaccessible from the outside, public documents in Mozambique and Cape Verde instances are to be available to local universities.
Currently, both repositories are deployed and functional, although they are still in an initial phase, but the usage of these instances will be intensified in the current phase. It is also planned to deploy repository instances in other PALOP countries, such as Angola.

3. Digital Preservation

The goal of digital preservation is "*to ensure continued access to digital materials... it refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change*" [18]. To provide the future use of information, digital preservation solutions must assure the access of data from multiple and heterogeneous systems (including unpredictable future systems). This requires preservation actions to maintain digital contents at the physical (bit stream preservation), logical and semantic levels. Assuring bit stream preservation is relatively straightforward, as replication strategies can be used to maintain the bit stream, while logical and semantic preservation require continuous action to assure that digital contents are still understandable. In e-Government solutions, trust is one of the qualities consumers of respective services value most [8], it must thus be at the centre of every such systems design.
According to the United Nations E-Government Survey 2010 [21] many African countries are, for various reasons, at the beginning of providing e-Government services to their citizens. With respect to digital preservation readiness this is not necessarily a drawback since African countries are thus able to build on a wealth of experience other countries made in the last couple of years [19]. Records management systems have been implemented and put into operation without taking the legal obligation to long-term preserve the records into consideration. As a consequence, records must be exported from live systems, converted into some dedicated compound format, and subsequently put into a separate repository intended for long-term preservation. Furthermore, in most scenarios used and later-on archived formats have never been harmonized, leaving the archive with a myriad of different file formats to cope with. Last, but not least, interfaces between commercial systems deployed as well as proprietary storage technologies render the integration of long-term preservation solutions more complex.
On the other hand, (open source) repositories deployed today, such as EPrints [15] or RODA[3], already have preservation functionality built in and/or support plug-ins to fulfil such tasks, with other systems currently integrating respective components.

3.1 The OAIS model and TRAC

Most archival information systems follow the generic standard, the Open Archival Information System (OAIS) reference model. It was published in 2002 and has proven to be a very useful high-level reference model, describing participants, roles and responsibilities as well as the exchange of information. The OAIS was adopted as a standard and registered under ISO 14721:2003 [20]. It does not recommend or specify any particular implementation but identifies and defines certain terms and components that might be

[3] http://roda.di.uminho.pt

involved in an archival information system. The OAIS defines *Producer* and *Consumer* as the main stakeholders and *Ingest*, *Access*, *Data Management*, *Storage*, *Management*, and *Preservation Planning* as the six main functional entities. Since institutions started to declare their repositories as 'trustworthy' or 'trusted digital repositories', concrete checklists such as Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) have been established; checklists against repositories can be audited [2]. The checklist is currently being standardized under ISO 16363.

3.2 Preservation Strategies

In recent years several different ways of preserving a digital collection have been evaluated, developed, and deployed. Most research on actual solutions for logical preservation is focused on two prevailing preservation strategies – *migration* and *emulation*.

Today, migration is the most common preservation strategy for reasons as the comparatively small initial effort necessary and price in terms of cost per object. Migration refers to updating a digital object by converting it from one hardware or software generation to another generation or representation. The main goal of migration is to transform to a more suitable or stable representation which is better suited for long-term access. The main challenge when migrating from one format to another is to only transform the logical representation, ensure authenticity and consistency and preserve all essential characteristics. The Council of Library and Information Resources investigated several migration projects and identified different kinds of risks [13]. As migration has to be performed continuously, i.e. at all times a file format is considered to be at risk, Jeff Rothenberg points out six problems of a migration approach [14]: (1) Labour intensive, (2) Time consuming, (3) Expensive, (4) Error prone, (5) Risky and (6) Non-scalable.

In contrast to migration, emulation operates on the environment of the object rather than the object itself. Emulation denotes the recreation of functionality of systems (software and/or hardware) which is needed to render, access, or edit a certain document to overcome technological obsolescence. In digital preservation this means mostly the emulation of a certain version of a software system needed to access a file in an obsolete version or format. Emulation is considered the better means for preserving complex objects as the more complex digital objects get, the more loss can occur in the migration process, i.e. the more significant properties can get lost. While having the advantage of leaving the original file untouched without modifications, writing an emulator is very complex and time-consuming. Additionally, there may be a point in the future where users no longer know how to interact with today's applications such as a certain word processing software.

3.3 Preservation Planning

As discussed above different strategies such as migration and emulation come into question for preserving digital objects. However, the decision for a specific tool e.g. for format migration or an emulator, as well as appropriate parameter settings for these tools, is very complex. The process of evaluating potential solutions against specific requirements and building a plan for preserving a given set of objects is called preservation planning.

The Planets Preservation Planning approach [16] allows the assessment of all kinds of preservation actions against individual requirements and the selection of the most suitable solution. It enforces the explicit definition of preservation requirements and supports the appropriate documentation and evaluation by assisting in the process of running preservation experiments. The result is a thoroughly defined, accountable preservation plan fitting the Develop Packaging Designs & Migration Plans functionality in OAIS [20].

3.4 Integrating DP in Repository Systems

Digital preservation is now supported by a wide variety of tools, each with their own distinctive interfaces. We are already seeing preservation tools that 'bundle' other tools to provide a specified workflow, e.g. File Information Tool Set (FITS)[4] , and the emergence of preservation systems such as RODA[5] [12] and Archivematica[6], that seek to manage complexity via single management interfaces. These aggregated tools and systems are now starting to integrate preservation support with the places where most new digital content is currently being deposited, stored and accessed, in the institutional repositories. An example of integrating preservation solutions with repository systems is offered e.g. by EPrints[7] [15]. The workflow for this consists of basically five stages, namely (1) identification of objects (2) characterization of their properties, (3) risk assessment, (4) preservation planning, and (5) applying preservation actions. To implement this workflow for EPrints digital repositories, KeepIt and its predecessor JISC Preserv projects have been applying tools for preservation workflow produced by the EU FP6 Planets[8] project: DROID for Format identification, XCL for Characterisation, and PRONOM and Plato for Risk assessment. KeepIt is integrating these tools into EPrints through a series of plug-ins that provide access to the tools.Similar types of integration are currently under way for other repository systems, particularly Fedora-based systems such as RODA in the scope of the EU FP7 project SCAPE[9] [5].

4. Recommendations for setting up and deploying archive solutions

While hard to summarize in just a few paragraphs, and thus necessarily incomplete, a few key lessons can be learned from e-Government and digital preservation initiatives that have been put in place so far:

(1) Identify need for DP: e-Government solutions only make the government's actions accountable and understandable to its citizens when they consider the whole document life-cycle – from creation, to archiving and potential re-use. Digital preservation must thus be considered from the early start. Adding the archiving component a posteriori often results in unnecessarily complex interfaces between both systems.

(2) Develop policies: Rather than optimizing individual services for mere task fulfilment, but then ending up with unpreservable objects or a myriad of formats, identifying standards results in more homogeneous collections, documentation of important aspects already as part of workflow design helps assuring authenticity of objects preserved.

(3) Plan the whole object life-cycle: *Is the long-term archive the final destination of the records?* This is an important question that has to be answered. *Or, do we need the objects regularly for being processed?* Two different philosophies very often come into conflict here. Most archival systems only support either of them.

(4) Design an integrated, modular, open solution: Designing DP services as an integral solution using clear (open) interfaces – as opposed to adding it "on top"/ in between existing complex, commercial software components with unclear interfaces, eases traceability and maintainability of systems. When data from a production system has to be exported, converted to a mutually agreed form, transferred to the archive and ingested, integrity is very hard to watch and guarantee. Apart from mere system complexity, all

[4] http://code.google.com/p/fits
[5] http://roda.di.uminho.pt
[6] http://archivematica.org
[7] http://www.eprints.org
[8] http://www.planets-project.eu
[9] http://www.scape-project.eu

parties, producer and archive, have to ensure that no errors have been introduced during transmission or storage.

(5) Consider organizational aspects: Introducing digital long-term archives into e-Government solutions not only poses technical challenges, it also demands assessment of the current organization. Digital preservation will partially overlap with existing workflows and organizational structures, whereas other aspects require new skills, staff training, and ultimately organizational structures.

(6) Small is beautiful: Digital repositories currently in operation are often large, complex and proprietary. Organisations try to buy a sense of security in the form of warranty (sometimes in peril of unmet requirements unique to certain organisations). However, as a start, slim systems with a modular architecture can be set-up in many cases directly by the organization. Open source repositories may in many cases fulfil many (most?) of the requirements, with the advantage of keeping the knowledge in-house and not being dependent on any single software vendors, even though IT support will frequently be sourced externally. This provides an opportunity to deploy preservable e-government processes with a low entry barrier.

5. Conclusions

European countries have been very active implementing e-Government solutions in the last couple years [19]. However, in very few cases the entire document life-cycle has been considered right from the start of these initiatives, covering its creation to its deposit into a long-term archive, and its possible re-use. Today, many different proprietary records management solutions are in operation in government departments making it difficult to find common ways for delivering to archives – no standard interfaces existed between systems at implementation time. Furthermore, policies have not been adapted to the new situation of born-digital governmental records which, in most cases, are a compound of files held in different file formats. This, ultimately, leaves archives with all possible file formats used in government departments, from simple text files to proprietary formats.

African countries may have lagged behind in the e-Government wave but can now leverage lessons learned of their European colleagues in this area. We see great potentials for African countries, currently in the process of implementing e-Government solutions since they don't have to deal with such legacy. Implementing repositories with digital preservation in mind makes it easier to actually have the solutions in place required for this task and makes it easier for archives to comply with their legal obligations. A solid requirements and risk analysis is a prerequisite for successful preservation. The sooner risk analysis is performed in the entire document life-cycle the less challenging is the task of long-term preservation.

Current initiatives: With a lot of effort having been directed to the preservation of document oriented objects, research projects in Europe are now focussing on a series of new challenges including the deployment of entire preservation workflows as in the ENSURE project. SCAPE, as already introduced above, is focusing on scalability issues, both wrt. massive scale deployment on cluster infrastructure, as well as support for automated quality assurance, monitoring, etc. TIMBUS[10], on the other hand, is addressing the preservation of entire business processes, rather than individual, static objects, identifying the aspects of processes and their context to be documented as well as means to verify re-invoked processes. This is an essential task specifically in e-Government settings when aspects of the process are essential. Details on these and other research activities in the area of digital preservation can be found in a project survey [22] recalling both the past as well as pointing to current and future trends.

[10] http://timbusproject.net

Acknowledgements

This work was partially supported by the projects SCAPE and TIMBUS. Both projects are co-funded by European Community under the 7th FP. SCAPE Grant Agreement number 270137; TIMBUS Grant Agreement number 269940.

References

[1] Samuels, H.W. (1986). "Who Controls the Past," American Archivist 49 (Spring 1986): 109, Quotation from George Orwell, 1984

[2] The Center for Research Libraries (CRL) and Online Computer Library Center, Inc.(OCLC). 2007. Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Technical report.

[3] Jimerson, R. (2007). Archives for all: professional responsibility and social justice. Am Arch 70:252–281

[4] Jimerson, R.. (2009). Archives power. Memory, accountability, and social justice, Chicago: SAA

[5] Edelstein, O., Factor, M., King, R., Risse, T., Salant, E., Taylor, P. (2011). Evolving Domains, Problems and Solutions for Long Term Digital Preservation. In: Proceedings of 8th International Conference on Preservation of Digital Objects (iPRES2011) pp 194-204 Singapore, November 1-4, 2011

[6] Netchaeva, I. (2002). E-Government and e-democracy: A comparison in the North and South. Gazette. The International Journal for Communication Studies, Vol. 64 No. 5 (2002) pp 467-477

[7] Kamar, N. and Ong'ondo, M., (2007). Impact of e-Government on Management and use of Government Information in Kenya, In: Proceedings of World Library and Information Congress: 73rd IFLA General Conference and Council. Durban, South Africa, 19-23 Aug 2007, At: http://archive.ifla.org/archive/IV/ifla73/papers/119-Kamar_Ongondo-en.pdf, Accessed: Dec. 9th, 2011

[8] Kelly, G., Mulgan, G., Muers, S. (2002). Creating Public Value: An analytical framework for public service reform. Discussion Paper. Cabinet Unit, London, UK, At: http://www.cabinetoffice.gov.uk/strategy/downloads/files/public_value2.pdf, Accessed: Dec. 1st, 2011

[9] Relyea, H.C. (2002). E-gov: Introduction and overview. Government Information Quarterly 19(1), 9

[10] Mbwete, G., Bhalalusesa R. (2011). Assessment of Public Services e-Government Initiaitves: A Case Study of Tanzania, Proceedings of IST-Africa 2011, Gabarone, Botswana.

[11] Ng'eno, E. J. (2010). Embracing E-Government in Service Delivery and Business to People Through Libraries: a Case for Kenya. Proceedings of IST-Africa 2010, Durban, South Africa.

[12] Ramalho, J. C., Ferreira, M. (2009). "RODA: A service-oriented repository to preserve authentic digital objects", Open Repositories. http://redmine.keep.pt/attachments/8/OR09-0.3.pdf

[13] Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., and Kenney, A. R. (2000) Risk management of digital information: A file format investigation. CLIR Report 93, CLIR.

[14] Rothenberg, J. (1999). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. Council on Library and Information Resources. http://www.clir.org/pubs/reports/rothenberg/contents.html

[15] Tarrant, D., Hitchcock, S., Carr, L., Kulovits, H. and Rauber, A. (2010). Connecting preservation planning and Plato with digital repository interfaces. In: 7th International Conference on Preservation of Digital Objects (iPRES2010), 19 - 24 September 2010 , Vienna, Austria.

[16] Becker, C, Kulovits, H, Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H. (2009). Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. In: International Journal on Digital Libraries (IJDL)

[17] Mnjama, N. (2010). Preservation and management of audiovisual archives in Botswana. In: African Journal of Library, Archives and Information Science. http://findarticles.com/p/articles/mi_7002/is_2_20/ai_n56577470 last accessed: 29. November 2011

[18] Jones, M. and Beagrie, N. (2008). Preservation Management of Digital Materials: A Handbook. London, UK: Digital Preservation Coalition

[19] Grönlund, Å. (2010). Ten years of e-government: the 'end of history' and new beginning. In Proceedings of the 9th IFIP WG 8.5 international conference on Electronic government, EGOV'10, pages 13–24, Berlin, Heidelberg, Springer-Verlag.

[20] ISO.(2003). Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003).

[21] United Nations E-Government Survey 2010. Leverating e-government at a time of financial and economic crisis. (2010), At: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan021888.pdf, Accessed: Nov. 25th, 2011

[22] Stephan Strodl, Petar Petrov, and Andreas Rauber. Research on digital preservation within projects co-funded by the european union in the ict programme. Technical report, Vienna University of Technology, May 2011. http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf