Open source versus commercial solutions for a long-term preservation in digital repositories

Andrea Fojtu

andrea.fojtu@ruk.cuni.cz

Charles University of Prague. Czech Republic

Abstract: The paper continues the comparison (already presented at the 16th International Seminar CASLIN 2009 - Institutional on-line repositories and Open Access) of substantial characteristics of today's available, well-known open source and commercial solutions for a long-term preservation of digital documents in digital repositories. At the beginning two major postulates prevailed: there would be at least one SW solution (open source or commercial) that would be suitable for dealing with the long-term preservation and comply to all the chosen criteria. The second one is related to a better performance of open source systems because of a widespread developer and user community. The reality was quite disparate from the formulated hypotheses. The results of the comparison led to several conclusions: development in the field of repositories for a long-term preservation is still in its infancy and more strengths has to be applied; and however fine the system is, it is not a redemption and human factors, financial resources, institutional policies plus risk management (including testing, auditing and certification) play a very important role and need to be taken into account.

Keywords: open source repository, commercial repository, digital repository, long-term preservation, digital curation OAIS model

1 An initial point

An impulse for the survey of open source versus commercial solutions for a long-term preservation was the Repository Software Survey, conducted by JISC (March 2009) and focused on comparison of different aspects [3]. The most important of them were: a, supported formats; b, thumbnails; c, user interface functions; d, advance search; e, browsing; f, classification / subject headings; g, user authentication; h, statistics; i, SW platforms, OS, scripting languages; j, metadata and k, interoperability.

It is noteworthy that in the above-mentioned comparison, functionalities of SW for repositories turned out be very even.

Notwithstanding, would be it similar when focusing on the long-term preservation?

2 Long-term preservation

The importance of the digital preservation may be corroborated by a Rothenberg's famous saying: "the digital information lasts forever or five years, whichever comes first" [2].

In the Anglo-American information sources we mostly come across terms like digital preservation, (long-term) preservation of digital objects or digital curation.

Digital preservation is the series of actions and interventions required to ensure reliable access to authentic digital objects for as long as needed.

The common terminology and conceptual framework for all projects dealing with the long-term preservation of digital documents is the OAIS model (the Open Archival Information System). It defines "*an archive, consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a designated community [...] and for 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*" [1].

3 Comparison & Results

In order it was possible to collate the open source versus commercial solutions (the most often implemented in today's information institutions), SW from the aforementioned study [2] were chosen: a, open source SW: DSpace, Fedora, EPrints and Research-Output Repository Platform; b, commercial SW: CONTENTdm, Digital Commons, Digitool, Equella, intraLibrary, Open Repository and Vital. Three relatively "new" systems were added to the evaluation, namely IBM Dias, Tessella SDB, Ex Libris Rosetta.

The criteria were prevalently focused on the long-term preservation: a, existence of OAIS model implementation; b, a wide range of supported formats; c, an open architecture for other applications and plug-ins; d, internal tools for format change (e.g. emulation, migration); e, a SW platform and HW (in)dependence; f, administrators' functions; and g, services.

At the beginning two major postulates prevailed: there would be at least one SW solution (open source or commercial) that would be suitable for dealing with the long-term preservation and comply to al of the chosen

criteria. The second one is related to a better performance of open source systems because of a widespread developer and user community.

The reality was quite disparate from the formulated hypotheses. It was biased by the very form of this survey - theoretical level (based on the search in presentations, articles, papers on the Internet). It was a very complicated and partially misrepresenting process. Moreover, information changes on the constant basis.

The comparison demonstrated that none of the SW solution complied to all the given criteria, however both commercial and open source SWs could be found in the promising group of the long-term preservation representatives.

A better performance of open source systems because of a widespread developer and user community was not confirmed as well. The fact is, only one open source solution, namely Fedora, complied to the majority of criteria.

Fedora's compliance: OAIS model implementation, METS, not PREMIS, open standard, OS and HW independence; dependence on PC – Midrange server, SIP as a 'compound digital object,' nonexistence of migration and emulation tools, indexing for full-text search. Not known: limits for (a bulk) ingest, ingest scheduler, versioning of digital documents and statistics.

Commercial counterparts showed better results.

SDB and Rosetta's compliance: OAIS support, METS (in case of SDB export to METS is possible), PREMIS, open standard, OS, SW and HW platform independence; SIP as a logical entity, versioning of digital objects, ingest scheduler, web archiving, statistics, indexing for full-text search, browsing, support in Czech Republic and no limit for (a bulk) ingest.

4 Conclusions

The results of the poster survey lead to two possible conclusions. **The first has to do with the very development in the field of repositories for a long-term preservation is still in its infancy and more strengths has to be applied. The latter one has to do with the fact that however fine the system is, it is not a redemption and human factors, financial resources, institutional policies plus risk management (including testing, auditing and certification) play a very important role and need to be taken into account.**

1 ISO 14721:2003, *Space data and information transfer systems -- Open archival information system -- Reference model.*

2 ROTHENBERG, J. *Long-term Preservation of Digital Information: Challenges and Possible Technical Solutions* [online]. NEDLIB talk, December 2000 [cit. 2009-04-05]. Avialable at: <http://nedlib.kb.nl/workshop/rothenberg_kb.pdf>.

3 *Repository Software Survey* [online]. JISC RepositoryNet, March 2009 [cit. 2009-04-05]. Available at: <http://www.rsp.ac.uk/software/surveyresults>.