Evaluation Report of the Project “Long-term access – from prototype to tools in operation”

Herbert Van de Sompel
Research Library
Los Alamos National Laboratory

Project Background

The Electronic Publishing Centre of the Uppsala University Library, led by Eva Muller, conducted the Project “Long-term access – from prototype to tools in operation”. The Project is a continuation of the SVEP Project, and is aligned with both the OpenAccess.se and the Digital Deposit programs of the National Library of Sweden that funded the Project:

- OpenAccess.se is generally concerned with facilitating access to works produced by researchers, teachers and students at Swedish universities and university colleges. One strand of OpenAccess.se work is aimed at providing long-term, secure access to these digital scholarly documents.
- The Digital Deposit project investigates technical solutions to support deposit of fixed and finite digital documents, in the context of the anticipated extension of the Swedish Legal Deposit legislation to also include digital materials.

Eva Muller left her position at Uppsala University Library at the end of January 2007 to lead the new Digital Library Department at the National Library of Sweden. Digital deposit is one of the primary tasks of the department. In her new position, Eva Muller will lead all efforts with this respect, including the creation of tools for receiving and validating digital objects deposited to the National Library.

Project Evaluation Background

In order to be able to prepare this report, the reviewer was provided with access to Project materials made available on the Project Wiki, a version of the original Project proposal, and an interim Project report containing suggestions for activities to be conducted with the remaining Project budget. Also, on December 13th 2006, the reviewer paid a full-day visit to the Electronic Publishing Centre, meeting with the Project Team, to gain detailed insights in the Project results, thus far. On December 14th and 15th 2006, the reviewer participated in the Seminar “Long term access to digital resources: current strategies and tools” in which the Project was discussed in the broader context of the (future) Digital Deposit mission of the National Library of Sweden, and the related digital preservation challenges. Representatives from other Swedish cultural heritage institutions and university libraries participated in this remarkably interesting and lively Seminar.
Because the approaches taken by the Digital Deposit program (as described by Stina Degerstedt in her presentation on December 14th) and the Project are closely aligned (as they should be), this report reflects on both, providing a mix of strategic and technical feedback.

Also, this report will focus on issues of potential concern only, providing food for thought when considering next steps and eventual reorientations. The report will refrain from qualifying the Project’s achievements along the lines of good-bad-ugly for the mere reason that it is obvious that the Project has delivered demonstrably solid work in a highly challenging problem space in which “The Real Truth”™ yet has to be established, and has done this within a relative short time period, with - judging by the size of the Project team - relatively modest means. It is also clear that the Project has significantly contributed to the objectives of the Digital Deposit project of the Swedish National Library, by investigating infrastructure to support collecting, cataloguing, archiving/preserving and accessing the scholarly communication subset of works that (will eventually) reside under Legal Digital Deposit.

The Project proposal states the following as Project objectives:

Within this continuation project the following prototypes will be developed in order to establish a finished tool that will be published under an open source code license:

a. Format and Data Environment Register  
b. Tool for packaging and transfer of objects to be preserved at KB based on standards for complex object format (METS, DIDL)  
c. Tool for checking and reception of objects at the local/national archives

In the presentations of Project results to the reviewer, the following were listed as deliverables:

- Related to (a) above:
  1. Format and Technical Environment Registry
- Related to (b) above:
  2. Contract tool  
  3. Metadata mapping tool  
  4. Complex object format tools (DIDL & METS)  
  5. SIP Packager
- Related to (c) above: none
- Unrelated to above, but resulting from predecessor projects and of crucial importance to the current project:
  6. Resolution Service

The reviewer notes that no deliverables were achieved for (c) above, and hence feedback will be restricted to (1) to (6). Furthermore, it should be noted that all feedback relates to work conducted up until the aforementioned seminar held in December 2006.
Project feedback

1. It is of crucial importance that efforts regarding Digital Legal Deposit and Digital Preservation become tightly coordinated among all major Swedish Cultural Heritage institutions. Such coordination is essential to achieve common solutions, and shared know-how (and maybe even some shared infrastructure). This will be beneficial to both the Cultural Heritage institutions themselves, and to their content suppliers and customers. Of specific importance is the sharing of outward facing procedures and (machine, user) interfaces for the acquisition of content from suppliers. While suggestions for such coordination and collaboration are commonly received in a most positive manner (because the benefits are obvious) they rarely move beyond the good intentions (because of the overhead induced by collaborations, and because of classic territorial sensitivities). Hopefully, the Legal Deposit and Digital Preservation challenges will provide enough incentives for a true collaboration to emerge.

2. Given the nature of the challenges involved in both Digital Legal Deposit and Digital Preservation, it is of crucial importance that the international landscape be continuously monitored for trends, procedures, standards, and technologies. Probably the most informative and rewarding manner to do so is to become actively involved in cutting-edge projects of strategic importance, and/or to act on the advisory boards of such projects. The role of the National Library of Sweden in the specification of the WARC file format for digital preservation is exemplary with this respect. Participation in international meetings, tough informative and important for networking, will rarely deliver insights as profound as can be obtained by actively engaging in leading international efforts.

3. Clear and broad communication about project goals, approaches and results should be considered essential. This is a proven approach to solicit peer commentary, but it also is an approach to have a constructive impact on the evolving international thinking regarding the Legal Deposit and Digital Preservation challenges. Telling from the nature of the in-person communications (both written and verbal) regarding the Project with the reviewer, it is clear that there is ample room for improvement with this respect. For example, it was well into the first day of the Seminar (after the full-day visit to the Uppsala University Library), that some light was shed on basic questions such “who is the user”, “what is the operating context”, “what is the source/target”, “what is the anticipated workflow, and how do the tools fit in”, etc. Being able to clearly communicate about the problem also fundamentally helps to think crisply about possible solutions.

4. The strong focus in the Project on the creation of high quality SIPs begs for some reflection. Most likely, this focus is inspired by a similar emphasis in the Digital Deposit effort. Without any doubt it is a noble goal, which, if achieved, could dramatically reduce the ingestion-related efforts at the end of the National Library.
However, related international projects suggest that achieving this goal is not realistic in many cases:

a. In the context of an NDIIP project, the reviewer has pioneered an approach for the standards-based transfer of high quality SIPs between a source repository and its mirror(s) \(^1\). The sponsors of the research (Library of Congress) have reacted very positively to the proposed approach, but there have been clear signals that it can only be deployed by a limited group of technically skilled source repositories. More specifically, representatives from the Prospero effort indicated that probably less than 10% of their content providers could consider deploying the approach.

b. The following email sent to the digipres@ala.org list expresses caution regarding what one can expect of submitters of SIPs:

```
Subject: [digipres] Autenticity and integrity of files in ingest
From: Karo Salminen <karo.salminen@helsinki.fi>
Date: Fri, 23 Mar 2007 13:23:53 +0200
To: digipres@ala.org

Hello all,

For short: when it comes to submitting or ingesting material to repositories responsible for long-term digital preservation, has anyone seen plans or even real implementations of a workflow or a tool, which would actually check the authenticity of the submitter in addition to the integrity of the received data?

Let’s take a look at PREMIS* for example. It provides semantic elements for storing the digital signature information. It would be really useful, if we saved the digital signature of the submitter in addition to repository’s own signature (stating that the stored data is authentic), because then it would be possible to check the authenticity of the submitted material afterwards.

We could, of course, identify and authenticate the submitter using for example login and password, and submitter could state that the SIP he/she submitted is authentic by clicking a submit button. But if we want to use digital signatures, the challenge is to use them in ingest in a smooth manner, making the process as straightforward and simple as possible for submitters.

We cannot assume that submitters would create, let’s say, a METS-formatted SIP in a zip file for us including their signature in GPG format. It would be the easiest way (and acceptable) for them to use a web browser interface for submitting files and adding some descriptive information about the files. Can you think any way to include digital signing of the material to that?

Some repository software, Fedora coming first to my mind, support checking and saving the integrity information of the submitted material, but only if the integrity checksum is provided to the software by the submitter. We cannot presume that the submitters would do even that, i.e. generate checksums for files they’re submitting.

* http://www.loc.gov/standards/premis/

Karo Salminen
System Analyst
National Library of Finland
```

5. Related to (4) above, it is regrettable that no survey has been conducted to validate the hypothesis that, given appropriate tools, potential submitters will in effect be

\(^1\) See http://dx.doi.org/10.1045/june2005-bekaert
willing to invest time and energy in using the tools and as such to deliver high quality SIPs to the National Library. It seems essential to conduct such a survey when considering a decision to further develop the existing toolset and turn them into a robust, deployable package.

6. In a similar vain, it is regrettable (and clearly a potential subject of criticism within Sweden) that, thus far, no attempts were made to test the tools in connection with a variety of heterogeneous repository systems (i.e. beyond DiVA repositories).

7. In any further development of the toolset, significant attention should be given to a data model for the representation of the object that is being packaged in the SIP. This notion is remarkably lacking from the current set-up which seems to be strongly metadata driven:
   
a. The Project delivers a metadata mapping tool but does not deliver an object mapping tool that allows for the mapping between the data model as used internally within a source repository and the data model used to represent the same object for submission to the Deposit system. For example, without a well-defined data model for the representation of objects in the SIP:
      i. How will e.g. structural information known to the source repository be expressed in the SIP?
      ii. How will the tools deal with objects that originate from source repositories that are not DiVA?
   
b. If future work will continue building on METS, efforts will have to be made to create a METS profile for the Swedish SIP application, to compensate for the lack of data model in METS. The Swedish METS SIP profile could fulfill the role of a data model, and then would allow for an unambiguous mapping from the internal data model of a source repository to the data model used for SIP submission.
   
c. The Project’s efforts regarding MPEG-21 DID/DIDL seem to be rather shallow. This is regrettable, given the ISO-standard status of DID/DIDL, the existence of a real data model for the expression of compound digital objects provided by DID, and the increased interest from other national libraries (the reviewer is aware of ongoing efforts in the Netherlands and Denmark) in the use of DID/DIDL. The reviewer admits to being biased with this respect, having advocated the use of MPEG-21 DID/DIDL, and having been involved in its standardization. At the same time, the reviewer’s preference is based on solid arguments resulting from detailed investigations of competing compound object representation approaches.
   
d. It is also interesting to note that no time was spent investigating the potential of the IMS/CP packaging approach. Especially given the preference of delivering zipped-up SIPs (a common practice in IMS/CP) this is surprising.

8. It also comes as a surprise that no real perspective is offered by the Project on the actual transfer mechanisms for the SIPs: is SIP transfer going to be push-based, pull-based, driven by automatic protocol interaction (e.g. OAI-PMH, RSS, SiteMaps, .), alert-driven, manually launched, … It seems that a perspective on this aspect of the problem is relevant because it can affect the choice of a SIP format (for example,
OAI-PMH could only deal with XML-based SIPs). And it further emphasizes the need for a data model for representing the object in a SIP because, given a solid data model, a variety of serializations can typically be defined.

9. Significant additional thought must be given to the emerging policy of assigning new identifiers for objects (urn:nbn) that already carry identifiers, and especially to the impact of this policy in the realms of e.g. discovery and citation. Under the evolving approach, a rather common scholarly object will have 3 identifiers: the identifier it received at publication time (say a DOI), an identifier it was assigned by the institutional repository in which it was self-archived post publication, and an identifier (urn:nbn) it was assigned by the National Library when it was being ingested for Deposit. Technically, there is nothing fundamentally wrong with this approach, but it does seem essential that the correspondence between these identifiers is recorded by the deposit system, that the identifier resolver supports the resolution of all these identifiers, and actually reveals their correspondence during the resolution process. The identifier resolver should also be extended to support a variety of service requests pertaining to identified objects (get metadata, get representation, get digest, get alternate identifiers, get copy from National Library, get local repository copy, etc.), instead of the mere redirect that is supported currently.

10. Given the long-term perspective that the Project and the Deposit efforts necessarily must have, the introduction of abstractions and “timeless” architectural design should be given consideration. Abstractions have the significant advantage that they can be instantiated using a variety of technological approaches at any given moment in time, but, even more important, different technologies can implement the same abstractions as time goes by. Examples where abstractions can come in handy include the aforementioned data model for the representation of the object in a SIP, transfer mechanisms for SIPs, the query interface for the identifier resolver, and the query interface for the format registry. Regarding both these query interfaces, the NISO z39-88-2004 OpenURL Framework Standard provides an interesting abstract platform to convey requests centered on an identified object (Referent) and an identified service request (ServiceType).

11. Difficult choices need to be made regarding the format registry deployment: one per source repository, one per country, an international one, etc. In this case, the reviewer would recommend a pragmatic approach combining the deployment of a rather straightforward format registry at the National Library (in concert with other Swedish heritage institutions) with the continuous monitoring and influencing of international efforts with this respect. It seems a safe bet that crosswalks between solid format identification approaches should be possible. Maybe efforts regarding format identification at the level of source repositories can be leveraged, but the National Library probably is better off trusting its own assessments regarding the format of datastreams then relying on those of source repositories that may not always be equally well administered.
12. Some consideration should be given to semantic typing of objects. This is a rather recent area of interest that may turn out to be crucial when creating meaningful applications on the basis of large collections of stored objects.