One process to rule them all? The role of a repository platform in the management of digitized cultural heritage at the National Library of Finland

Jyrki Ilva and Esa-Pekka Keskitalo, National Library of Finland
E-mail addresses: jyrki.ilva@helsinki.fi and esa-pekka.keskitalo@helsinki.fi

Although repositories are generally associated with green open access and research publications, they are commonly used for the storing and dissemination of other kinds of materials as well. Many organizations dealing with scholarly publications also have extensive collections of cultural heritage materials. The National Library of Finland is no exception.

In this paper we look at some of the strengths and limitations of DSpace repository platform from the point of view of digitized cultural heritage collections. We argue that the repository should not be seen as a standalone system, but as part of a larger organization-wide system infrastructure. When digitization is being done at a mass scale, it is essential to have well-planned processes for the dissemination of metadata and digital files from one system to another. We highlight some of the challenges the National Library is facing on its road towards better interoperability and integration.

1. Cultural heritage collections at the National Library of Finland

Like most of the national libraries across the world, the National Library of Finland has rich collections of cultural heritage materials, from the Middle Ages to the present day. As a legal deposit library it has near-complete collections of Finnish books, newspapers and audio recordings. It is also harvesting Finnish web materials for the Finnish Web Archive.

The Library has been actively digitizing its collections since the late 1990s, and has an in-house digitization centre. As a general policy the library aims to publish all of the materials it digitizes as openly as possible. However, due to Finnish copyright law, most of the more recent materials may be used only within the premises of the legal deposit libraries.

2. Cultural heritage collections on a repository platform?

DSpace has many basic functions that it performs (at least) reasonably well. This is something that should not be taken for granted with all of the digital asset management systems available even today. Although DSpace was originally conceived for a relatively specific use case, it is generic enough to be used for various purposes. It can be connected to other systems via technical interfaces, and as an open source platform, can be modified with relative ease.

The National Library originally ended up using DSpace as a replacement for a proprietary platform that had not fullfilled expectations. (The other option on the table at the time, building on Fedora, was dismissed due to time restrictions.) Our first public DSpace instance, *Doria* (http://www.doria.fi), was launched in February 2007. In addition to the collections of the National Library, it also housed the institutional repositories of several customer organizations.[1]

Even later on, the National Library examined the possibilities of using another proprietary system that had built-in support for the METS format files produced by the digitization process. There were also plans to expand an existing system built for the digitized newspapers and journals to manage other kinds of digitized materials as well. However, this did not prove to be viable, and the use of DSpace for the materials that are not newspapers or journals turned out to be the best option. Although DSpace is not perfectly suited for all kinds of cultural heritage materials, it is good enough for most of them.

3. Repository as a part of a larger system architecture

The cultural heritage materials being digitized are often very heterogeneous and cannot be processed in one generic way. Obviously, it would be preferable to have one general process for all of the materials, but in practice a lot of adjustments need to be made, and short-term ad hoc solutions are sometimes hard to avoid.

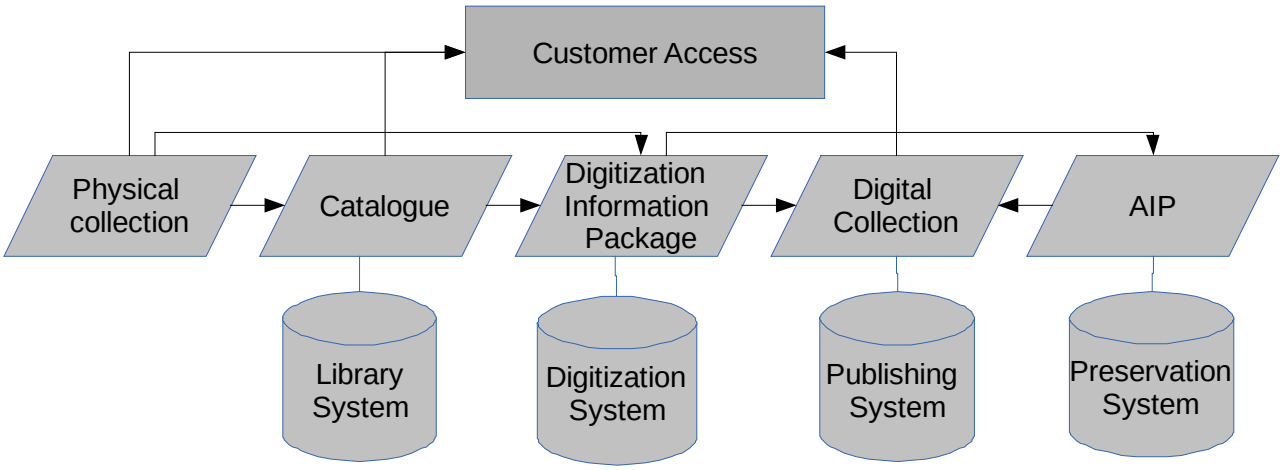
It would not make sense to create a new platform for each new collection. Any new system would produce both initial adoption costs and long-term management costs. In our case this would mean that we would eventually have hundreds of different collections in many different legacy systems built for projects that no longer have any funding. Of course, this would be a nightmare scenario for the IT management team of any organization.

---

[1] Currently the National Library has six public DSpace instances and is providing repository services for 38 organizations.

It seems to make much more sense to build a generic system or a small number of generic systems that are flexible enough to accommodate different kinds of materials. Of course, compromises must be made between collection-specific needs and the general manageability of the infrastructure. There should also be a general process for the handling of the digitized materials - although it is likely to be modified to suit each type of materials, many of the basic tasks and procedures will remain the same.

Although repository managers like us may sometimes have a repository-centric view of the organization-wide infrastructure, the repository is actually just one piece in a larger puzzle. The following diagram provides a rough picture of the role of different systems in the digitizing process at the National Library:



The repository acts as the publishing system, which is used for storing the digital items and their metadata in order to provide access to them. However, it is dependent on other systems which all have their own roles in the process.

It is quite common that the metadata for digitized cultural heritage works can be obtained from other sources, mainly library catalogues. At the National Library the production of metadata and digitization are usually done in separate units within the library. The processes are mainly designed for mass digitization: the digital items and their metadata are harvested to the repository from other systems, and they are converted from other formats (METS, MARC21) to the ingest format used by DSpace. The built-in manual submission tools of the repository are not used except when there is a need to import a very small number of items with no connection to larger processes.

It is not enough to build technical interfaces to connect the systems involved in the process. There is also a strong need for developing co-operation and mutual understanding between humans. This is at times somewhat challenging at the National Library, since the units involved are all in different geographical locations and under different leadership. To make the library-wide processes work seamlessly, all parties will need to share the same goals and work together to achieve them.

4. Some special challenges of cultural heritage collections

The following is a list of some repository-related challenges that we have run into while working with the management of digitized cultural heritage data. Many of the challenges cannot be solved solely within one system ; to properly solve them, one needs to think of all of the systems and workflows used in the cataloguing, digitization and publishing processes.

4.1. Increasing needs for interoperability between systems

Ephemera (http://www.doria.fi/handle/10024/85119), a collection of about 5,000 mass-digitized small publications from the 19th and early 20th century, is one of the collections that has benefitted from automated connections between the different IT systems within the library. The publications are catalogued into a bibliographic database, after which the MARC21 metadata is harvested into the digitization system. Files created, and the metadata, are transferred in METS format from the digitization system and converted to the ingest format used by DSpace.

If there is something wrong with the data, the corrections need to be made to the METS data. It is then re-sent from the digitization system to the repository, and the new record automatically replaces the old version. At this point this process works only between these two systems and still requires some human intervention, but in an ideal world all of the systems from the bibliographical database to the digitization system, to the repository and to the long-term preservation system would be connected and the changes that have been made in the metadata or full-text files could be communicated between them in real time.

We are not quite there yet, but still it would be of outmost importance to establish a workflow with which the data stored in different systems can be managed properly. In addition, there has been some concern over the loss of information in conversions from one format to another, most notably from MARC21 to (Qualified) Dublin Core. From the point of view of

DSpace and its users this is not a major issue, but must be taken into account in the planning of the total architecture.

Another issue that is still under consideration is which version of the data will be used in the national and international discovery systems, including the national *Finna* discovery interface (http://www.finna.fi) and *Europeana* (http://www.europeana.eu/). The persistent URN links to the full-text publications are being added to the bibliographic databases as well, which means that they could be also used as a source for OAI-PMH harvesting instead of the repository.

4.2. Unusual metadata fields

The metadata of cultural heritage objects may be radically different from the metadata of present-day publications. The items may contain unusual metadata fields, or they may be missing information that is usually considered essential, like exact publication dates, author names or even clearly defined titles.

A good example of this is provided by *Fragmenta Membranea* (http://fragmenta.kansalliskirjasto.fi), which is a collection of about 1,500 parchment fragments, containing most of what survives of the medieval literature circulated in Finland before the Reformation. The metadata of the collection was compiled by a research group specializing on the identification of medieval manuscripts. As the display of the data required a lot of modification to the user interface of DSpace, it was decided to create a dedicated DSpace instance for it instead of using Doria.

One of the challenges with Fragmenta Membranea was the use of inexact dating information in Dublin Core metadata fields and DSpace indices. Usually the publication date of a repository object is a day, a month or at least a year, but in this case we had to figure out how to express datings like “somewhere around the turn of the XIII century”.

4.3. Multilingual metadata

The user interface of Doria is available in three languages, Finnish, Swedish and English. Finnish and Swedish are both official languages in Finland, but the English language version can be understood by a far bigger potential audience. The library has also built a Russian language user interface to a new DSpace instance devoted to Fenno-Ugric materials that are being digitized in co-operation with Russian libraries.

On the other hand, most of the collections contain metadata only in one of these languages, usually Finnish. As some of the collections are of interest to international community as well, it would be beneficial to include metadata in multiple languages so that the information shown to a user would change according to the language of the user interface. As far as DSpace is concerned, this is not a major technical challenge, and has already been done to some extent in some Doria collections, including the *Dissertations of the Royal Academy of Turku* (http://www.doria.fi/handle/10024/50699).

However, multilinguality is a problem for some other parts of the library-wide process. If the metadata is imported from other systems, the translations should be kept in the same place with the master data in case it should be updated at some point in the future. Unfortunately, the library catalogues which are the most prominent source for metadata don't currently support the use of multilingual metadata. This means that the translations must be added later in the digitization process, which adds a layer of complexity.

4.4. Where is the master data?

The question of master data is crucial. Since both the metadata and digitized files are propagated from one system to another, there need to be clear rules about at which points the data is updated and at which points it is not. At the moment it is assumed that the main authoritative source for metadata is *Fennica* (http://fennica.linneanet.fi), the Finnish national bibliography, while the combination of metadata and digitized files is stored in METS files designed for long-term preservation purposes.

Fennica, however, has challenges as a master data source. First, it may contain only bibliographic metadata. Data regarding all other aspects of digital materials must reside elsewhere. Second, there is the multilinguality challenge described above. Third, synchronising changes in metadata is a complicated. Changes should be made first and foremost in Fennica. At the moment, both having the changes made and propagating them to the derivatives is burdensome.

4.5. Tools for display

Some of the image and pdf files produced in a digitization process are quite big in size, which means that they may not work very well even with modern computers with fast Internet connections. This is true for digitized books, manuscripts and historical maps, which need to be digitized in high resolution. The National Library has tried out several options to ease this problem, but is still working on it.

The in-house digitization process at the National Library produces METS files that are suitable for long-term preservation. These METS files contain all of the information generated in the digitization process, including the image files, OCR'd text, technical metadata and also the MARC XML metadata corresponding to the digitized work. Since DSpace cannot natively display METS files, the library has developed a separate HTML5-based METS viewer, which can be connected to DSpace records and used for the browsing of a work one page at a time. The OCR'd text is also available for searching.

In some collections the objects are available in three different formats: As a pdf file, as separate high-quality image files, and as part of a METS file which can be read following the link ("Viewer") to the METS viewer. The METS viewer works quite well, but it has been recognized that the number of different browsing options may appear confusing to the users.

4.6. Tools for annotation and correction

Digitized cultural heritage materials would often benefit from user-based annotation or correction of OCR'ed text, either by scholars or by the general public. The National Library has tried out crowdsourcing the correction of OCR'ed texts by using a computer game designed for this purpose, but the pilot was run separate from the repositories. The major challenge in this is how to map the results of crowdsourcing back to the data stored in the library infrastructure.

The National Library is currently working on a project funded by the Kone Foundation. It aims to provide annotation and correction tools for scholars of Fenno-Ugric linguistics. The library is digitizing Fenno-Ugric books and newspapers published in the Soviet Union in 1920s and 1930s in co-operation with two prominent Russian libraries. The original publications are of poor printing quality, and the obscure languages are not supported by current OCR software. Therefore the correction of the OCR results is expected to require a lot of work, especially as the aim is a near-perfect result for the purposes of linguistic research.

A special DSpace instance (*Fenno-ugrica* will be opened to public in May, 2013) has been launched for the project, which aims to create a large mass of linguistic data that can be used for data mining and computer analysis. The project has also produced an OCR correction tool, which makes it possible to edit the text contained in a pdf file. The image of each page is shown beside the OCR'd text, and each character in the text is mapped to its location on the page. The tool uses the same HTML5-based technical solutions that are also used in the METS viewer developed by the library.

Another example of the need for annotation tools is the *Literature Bank* (http://www.doria.fi/handle/10024/88083), a collection of nearly 600 classic 19th and early 20th century books, all of them out of copyright. The digitized books have been made available as pdf files in Doria. The OCR'ed text will be corrected using same tools as with the Fenno-Ugric materials. Information on the authors and references to scholarly treatments, contemporary reviews, etc. will be added.

4.7. Creating connections between resources

Cultural heritage data typically has lots of potential connections to other data stored elsewhere. Providing links between objects and data stored in different systems would in many cases improve the user experience and provide contextual background information for the objects.

A good example of this is the collection of Dissertations of the Royal Academy of Turku at Doria, which currently contains nearly 1,800 academic dissertations from the 17th, 18th and early 19th centuries. The person names mentioned in each record have been linked automatically to the biography of the person at the *Student Register 1640-1852* web site (http://www.helsinki.fi/ylioppilasmatrikkeli/) and vice versa.

Of course, some of the cultural heritage collections would also benefit from being connected to proper name authority service. The National Library is working on a national name authority service, but at this point it is still unclear whether this data will be used on the repository level.

4.8. Re-use of the data / metadata

It makes sense to license cultural heritage collections with out-of-copyright materials as open data. This will make it easier to utilize the metadata and the full-text publications in the creation of new services. These include web exhibitions, narratives and larger discovery services which may combine data from many different sources. Open data can be also used for data mining, although DSpace as a technical platform is not particularly well-suited to support this.

To ensure the widest possible dissemination for the materials, the National Library has decided to use the CC0 license for the metadata of its digital collections, and also for the out-of-copyright materials they contain.

5. The big question: what next for digitized newspapers?

The amount of cultural heritage data currently stored in Doria, Fragmenta Membranea and Fenno-ugrica is still quite small compared to *Historical Newspaper Library* (http://digi.kansalliskirjasto.fi), which contains all newspapers published in Finland up to year 1910 and most of the other periodicals as well. The service is run by the digitization centre of the National Library, and it contains millions of digitized pages. It is hugely popular among historians, genealogists and the general public.

The technical platform used by the Historical Newspaper Library is already showing its age and will need to be replaced within a few years, but the Library has not yet decided which route it will take. Will the new platform be based on open source or proprietary software, will the development be done as in-house work, as a community effort or will it be outsourced? One of the things we will keep an eye on is how well the next generation of repository software will be suited for this task - and for the needs of cultural heritage collections in general.