Submitted on: 13/08/2014

Rehousing digital heritage. Preservation on a very large scale

Tanja de Boer
Head Collection Care, National Library of the Netherlands
Tanja.deboer@KB.nl

Maarten van Schie
Business Information Manager, National Library of the Netherlands
Maarten.vanschie@KB.nl

Barbara Sierman
Digital Preservation Manager, National Library of the Netherlands
Barbara.sierman@KB.nl

Astrid van Wesenbeeck
Project leader, National Library of the Netherlands
Astrid.vanWeesenbeeck@KB.nl

Abstract

More than ten years ago the e-Depot of the National Library of the Netherlands (KB) became operational; one of the first long-term archives for international scientific publications, worldwide. Millions of publications, mostly e-journal articles in PDF format, were ingested, stored and made accessible.

For digital preservation and for scaling purposes the KB started in 2009 renewing its digital storage environment and building and equipping a new digital preservation system. A first release of the newly built system was dedicated to receive the millions of stored articles and to process the migration which would phase out the initial e-Depot. We could then start migrating collections to a new, better and future proof storage.

To move our digital collection, we needed to migrate these objects to the new hardware- and software environment.For KB, migrating millions of publications meant a major preservation action on an unprecedented scale. And one with possibly great risks for our digital heritage when not properly executed. It was not only a media migration. Almost 5 million publications had to be repackaged, which meant all files and metadata were checked and reformatted into new archival packages. We developed an automated workflow to process the material as well as processes for quality control and error handling.

The entire operation took 10 months and cost around €110.000. Almost 20 KB-staff members ensured a controlled and safe preservation process. Two members of staff were dedicated to monitoring the daily progress of migration and report on any errors. We set a fault tolerance for the migration of the files, and the project had to meet a strict deadline.

After careful preparation and testing the actual migration started in January 2012. We chose a two-step approach. The first step was migrating collections from the old e-Depot environment to temporary storage on a secure environment 'silent-cube'.[1] *In the second step, collections were transferred from temporary storage to the new digital preservation system. This two-step approach limited our dependency on the relatively poor processing capacity of the old e-Depot system.*

Migration was concluded in October 2012, with exception of the event metadata. At that point almost 5 million objects were successfully migrated through the two steps to their final new location and were instantly available to our patrons.

KB has greatly learned and profited from this project. We have gained experience in executing a complex preservation action on a large scale. And that will help us planning and executing future preservation actions, and also on more complex digital objects. We hope to share our lessons learned with fellow libraries as well as learn from their experience.

---

1. Introduction

The Koninklijke Bibliotheek, the National Library of the Netherlands (KB) aims to offer sustainable and reliable storage and access for digital publications from Dutch publishers, international publications about the Netherlands and from international academic publishers.

For this purpose the e-Depot was launched in 2003. The system was designed in conjunction with IBM. It served for almost 10 years and at the end of those 10 years, stored 18 million articles and books from various publishers.

Due to new digital archiving requirements and developments in the nature and number of born digital and digitised publications we intend to preserve, KB started to develop a new digital preservation system. The first release of this digital storage system, we call it "Digitaal Magazijn" (DM), took place in 2012.

It enabled us to transfer all stored content from the then current and now obsolete e-Depot-system to the new system. We started the project 'Migration', to move these 5 million objects, mainly academic journal articles, from one storage location, via a temporary station, to the new DM. The aim of this project was to do that in an controlled manner, compliant with the principles of digital preservation.

The first step, from e-Depot to temporary storage, started in January 2012 and was finished in September 2012. The second step, from the temporary storage to the new system via a repackaging process, started in May 2012 and was finished in October 2012. The project was wrapped up by the end of 2012.

The former e-Depot system

The e-Depot of the National Library became operational in 2003 as a long term storage system.It was one of the first systems for long term storage of digital publications of its kind. The system was composed out of standard IBM middleware components: Tivoli Storage Manager, Access Manager,

[1] The Silent Cubes system is the first storage system to be explicitly and categorically developed for the secure, long-term storage of permanent data.

Content Manager, WebSphere Application Server and DB2. The functional JAVA software was developed and maintained by IBM. The software solution was called DIAS (Digital Information Archival System). The archived files were primarily stored on a Plasmon optical storage system and secondary on tape on-site and a tape copy off-site.
The design of the DIAS system was based on the OAIS[2] reference model, or better ISO-14721:2012, a high-level reference model for the design of a digital archive. [3]

Several workflows were designed in the DIAS system to ingest different types of material: E- journals from Elsevier, a generic workflow for e-Journals of other publishers and a workflow for CD-ROMs.

The Elsevier journals bulk workflow was the first workflow to be built and was launched in 2003. After that, other publishers joined in. To have more control and flexibility when taking in new publishers the generic stream was implemented by IBM and the specifics for every publisher like FTP-schedules, normalisation of filenames, packaging and checks were implemented by the KB using a "pre-process". The pre-process was specified and developed by KB and added the flexibility to make fast adaptations to keep the ingest flow running despite the versatile nature of the material.

Digital objects

The majority of publications at that time consisted of articles from online journals mainly in PDF format, although other formats were allowed like TIFF, DOC, JPG, etc.. Most non PDF formats could be found in the supplemental files. CD-ROM publications were stored according to a different procedure, as a package with a full disk image of the installed software publication. Also some e-books and the material harvested from university repositories, mostly in PDF format, were stored.
All in all, more than 18 million publications were ingested and stored in our former e-Depot.

Processing in the former e-Depot was limited to e-journals and a small flow of e-Books. There was no possibility process websites nor was it possible to ingest digitised content. The e-Depot was not equipped yet, for handling the output of digitisations programmes. The KB had started digitising paper collections some 10 years ago. First we did small scale, mostly boutique projects learning the business of mass-digitisation. We currently have fine-tuned workflows in place for digitisation millions of pages (scanned images) per year. These projects are either publicly funded through our National Conservation Programme Metamorfoze, funded by public partners, or they are the result of public-private partnerships such as those with Google and ProQuest. The output of mass digitisation will be our major storage challenge for the years to come.

2. Aspects of the new DM

The need for a new system

New standards for digital preservation, new object types of born digital publications (like the web archive) and the surge of digitised material, lead to the conclusion that the e-Depot would not be suitable for scaling and expanding. Around 2009 KB started the development of a new system that should meet the requirements of contemporary born digital publications, and that would also be able to ingest and store vast amounts of digitised files. The new DM is OAIS-compatible, as was the former e-Depot.

Developing a new digital preservation system also implied inevitable that migration of our digital collections and their metadata. In 2010 the e-Depot of the KB had been in service for seven years

[2] Open Archival Information System - Wikipedia, the free encyclopedia

[3] For more information on the updated version of ISO 14721 see: http://digitalpreservation.nl/seeds/standards/oais-2012-update/

when the requirements for a new DM were developed and the data migration to the new DM or Digital Preservation system was prepared.

Migrating would pose quite a challenge because of the scale and heterogeneity of the archived material. In the new DM the design of the Archival Information Package (AIP) was changed to facilitate the storage of more meta data (amongst other reasons), including the descriptive meta data. Previously in the e-Depot system, basic (original) meta data were stored in the AIP and extended descriptive meta data in the catalogue system. This change in view had its consequences for the migration as the objects not only had to be moved but it should also be repackaged, as is the OAIS phrase.
In the DIAS system the information of the AIP was distributed over several places in the system. To create a new AIP for the DM these parts needed to be collected in order to create the AIP. The output of this repackaging process would be a new SIP (Submission Information Package) for the DM, in which several elements were combined.

The elements for the new SIP

The four elements that should lead to the new SIP are:
- The original Archival Information Package, a tar-file exported as DIP (Dissemination Information Package)
- bibliographical information as stored in the Catalogue of the KB (KB-MDO[4])
- technical information from the DIAS Content Manager (CM)
- Access information from Tivoli Access Manager (TAM)

The KB-MDO could contain more detailed descriptive information on the objects. Alos some errors like typo's in the title of a journal, would be corrected there and not in the original metadata. Therefore it was decided to include the KB-MDO metadata in the SIP for the new DM.

An extraction of the technical metadata was stored in the Content Manager. This information, which contained for example the file format information and the file size would be part of the new SIP. The fourth and last part consisted of the Access Information. Over the years the access information was added to each digital object, based on the publisher that delivered the journal. With some publishers we have an agreement to give restricted access to the articles, other journals are open access. As this is important information for every user, it is used in the retrieval system and described in the Tivoli Access Manager and should be included in the new SIP.

A flow was designed to retrieve this information and to check whether the information was correct. It was known that in the past some errors had occurred and we needed to take into account what the unavailability of an element could imply. For example, the absence of technical information in the Content Manager system was something the original DIAS system did not tolerate, so if this situation should occur, it was a sign that something was seriously wrong.
On the other hand, we knew that in the processes of synchronizing the DIAS system and KB-MDO, some errors occurred in the past. Absence of bibliographical information could be traced back to these errors and was as such acceptable for the migration. For this reason the check as well as the outcome of the check was documented, in order to be able to translate this in a later stage into relevant event metadata.

[4] MDO meaning metadata storage

Repackaging

It was important to design the repackaging process in a way that no information could get lost and that there was a clear auditable process, on the basis of which we could monitor deviations, create reports and event metadata. To support this requirement, we set up a separate "Administration database", in which each Archival Information Package from the DIAS system was represented in a separate record via the unique identifier, the National Bibliography Number (NBN)[5]. In this database we administered the status outcome of each activity on the information package for each record.

A specific workflow described the steps of collecting the four elements, the expected outcome of each step, the expected deviations and how to handle these. Another document described all possible combinations of situations that could occur (either the presence or the absence of one of the four elements) and highlighted the contradictions. These documents were intensively discussed with specialists in the KB who in present and past had dealt with the DIAS system, in order to avoid surprises. It also lead to one of our lessons learnt: had we recorded events better and in more detail in the past ten years, we might have had a better overview of what to expect.

Specifications

In the course of migration or repackaging bit level preservation had to be ensured, as well as the completeness of collection and files. Also we decided on independent internal monitoring or auditing the process

3. Issues and answers

Two main decisions had to be taken in the preparation phase of the project.

The migration had to be executed in two steps. Firstly all content was to be transferred to a temporary storage system (step 1). After that, the content was to be ingested into the new DM (step 2). A two-step migration helped us to manage two unsecure situations. Firstly at that point in time it was yet unsure what functionality would be available in the new DM. Secondly we didn't know yet at what time the DIAS system would be disabled due to contract termination. We wanted the objects transferred as soon as possible. By migrating all objects to the Silent Cube[6] we created a safe haven for our content. Other reasons to manage a two step migration will be outlined below.

Not all content would be transferred. It would sound reasonable and logical to transfer 100% of the digital preserved content, though the KB made some tough decisions which narrowed the scope of the project and rejected part of the content for the repackaging process. The so-called "no-content category" was defined, as amongst others test files and files that were corrupted in the starting phase of the e-Depot. We tend to call these exceptions 'no-content-files'. Their metadata were stored in the Migration Database, and we yet have to process them for provenance reasons.

Also excluded from migration were all articles from the publishing house Elsevier were excluded from the migration. The Elsevier articles made up about 50% of the e-Depot. Encouraged by Elsevier we decided not to migrate them but to reload them in a new and better format. The benefits of loading new objects were considered to accede the benefits of including all Elsevier articles in the Migration-project.[7]

---

[5] NBN: National Bibliography Number. http://en.wikipedia.org/wiki/National_Bibliography_Number

[6] FAST LTA Silent Cube Long-Term Storage **http://www.fast-lta.de/en/**

[7] The issue of not migrating Elsevier publications was presented at the 2012 IFLA conference: http://conference.ifla.org/past-wlic/2012/102-boer-en.pdf

Though we didn't transfer these so called no-content objects the project was required to register several metadata of these objects to create a record later showing the deleted objects for provenance reasons.

There were other reasons which led us to decide on the two-step approach. The time needed to execute step one gave us time to get step two ready. The DIAS system had a relatively low processing capacity and the two step method gave us some time to work around that.
It also gave us the opportunity to check and assemble the four elements needed to form the new SIP and prepare them before combining them in their new package.

4. Step one: from DIAS to the temporary storage

Step one of the migration process was called the Preparatory Process, consisting of several programmes and scripts that transferred objects from the e-Depot to temporary storage. The aim of this step was to create an auditable process to extract and check the four elements from the e-Depot system that would form the new AIP from the DIAS system. A secure temporary storage solution (silent cube) was used for storing the retrieved data and an administration database was used to keep track of the files, timestamps and statuses on the storage.

Retrieving large amounts of AIPs from the e-Depot

The main element for the new AIP was the old AIP, containing the publication, supplemental files, original metadata and table of contents of the AIP. The AIP was stored as a TAR file on the Optical Storage system. The TAR file could only be retrieved from the e-Depot using the default access interface, which is a website also used for public access. This interface, and the software and storage supporting it, were not optimized for large scale retrieval. The bottleneck was the optical storage solution for the e-Depot. This storage solution was chosen for the long retention time of decennia but the speed of access and throughput of files was slow compared to for example storage on hard disk. The files were stored on UDO optical disks,[8] and these disks were stored in a closed cabinet. When a file needed to be accessed, the specific disk was mechanically retrieved from the cabinet and placed in an UDO drive where the disk could be read. Because of the mechanical process involved, swapping disks took up around one minute.

To optimize the reading speed, files had to be read in the order in which they were stored, to minimize swapping of optical disks. The project team could do this with relative ease, as the identifier we used for retrieving, the NBN, was derived from a timestamp. Publications stored close to each other in time were also stored on the same optical disk. By numerically sorting the identifiers, retrieving was ten times faster compared to random retrieving. Going from roughly five seconds to half a second per publication.

Collecting the bibliographical metadata

The second important element to collect was the bibliographical metadata, which was stored in the database KB-MDO supporting the catalogue. The database supported OAI-PMH and SRU to harvest records. The project team used these services to make a harvest of all the bibliographic records and stored the xml files in the native Dublin Core eXtended format on a filesystem, indexed on NBN in the Administration Database.

The main reason for this separate metadata store was to reduce the project risks, as the current KB-MDO catalogue system was undergoing an upgrade and we needed to be sure the data could be used when we needed them in a stable format. The second reason was that versioning information of the publication was stored in the KB-MDO records but not indexed so it could not be retrieved directly.

[8] Ultra Density Optical (UDO) is an optical disc format designed for high-density storage of high-definition video and data.

Content Manager Database

The third element we retrieved was a selection of data from the DIAS Content Manager Database from the e-Depot. This contained technical meta data like the date of ingest of the object, AIP checksum, file size, file count, supplier etc. This data was exported from the database and stored as an xml file.

Access rights

The last element were the access rights of the publication. The access information was kept in a database and used by the Tivoli Access Manager. It contained detailed information on access, based on publisher and in some cases on journal level. Three levels of aces were possible: onsite, online and 'dark', meaning no access at present. The access level depends on the agreement with the publisher or on the journal title. A copy of this information was made and stored separately.

5. Step two: from temporary storage to new DM

The goals of the second step were checking, repackaging to a new AIP and storing the prepared publications in the new DM. The process was managed by software that checked the Administration database on which publications were prepared and ready for further processing. When the publication passed all the tests a message was sent to the ingest process with the location of the publication elements. This message was put on the queue of the ingest process, then the message queue was processed by multiple ingest pipelines.

The ingest process consisted of four main steps Checking , Repackaging, Storing and Manual Error handling.

Checking

The TAR file containing the publication and the checksum was recomputed and compared with the stored checksum. Then the Tar-ed AIP was unpacked and the files were checked against the Table of Contents file that was part of the AIP; a check was done on filename and file size. All XML files were validated: the Table of Contents file, Content Manager file, Access file, and Bibliographic metadata.

Repackaging

The second step consisted of combining and remoulding all XML metadata files in a so called "AIP manifest" using METS as metadata standard. In this process-step implicit information was made explicit by identifying the original metadata file in the old AIP, using business rules. During this process unique numbers were generated for the separate files, that had to be stored. The original identifier (NBN) identifying the whole publication was kept.

Error Handling

If an error occurred during checking or repackaging, the message was parked and handed over to a manual error handling process where errors could be analysed and resubmitted on the queue.

Storing

In the last step the AIP was stored, including the manifest. The process for storing consisted of the following steps. First the AIP-Manifest was imported. After a storage confirmation of all the files the AIP manifest was offered to archival storage, also a Silent Cube system, but with a head unit. The head unit is the portal to the silent cubes and includes a caching system where the files are stored for a short time before they are committed to the Archival Storage.

6. Results

Successful migration

Success was phrased as a 100% repackaging and ingest of the selected objects into the DM. This was the case. Of the 4.995.846 objects only 156 were not migrated. 134 publications had to be migrated a second time because of missing files, caused by an interruption in storage. This couldn't be fixed until January 2013 because of the necessary implementation off the delete procedure in the MDS-database. The processing of these NBN's was taken up by the business information manager. The other 22 publications couldn't be migrated for a variety of reasons, but in call cases this has been since fixed, or will be fixed in the future.

Time and money

The migration project was executed between January and December 2012 and cost around €110.000, which turned out to be about half of the expected costs. We spend less than €100.000 on personnel costs and a little over €14.000 in material costs.

Monitoring

Two of our staff, the digital preservation manager and the senior advisor for digital collection care, monitored the entire process. They participated in the design of the workflow and monitored the migration-logbook which was daily kept by the project team to follow all migration processes. They had frequent meetings with the project team and also executed a modest sample test. This confirmed that objects were successfully stored in DM. They drew up a separate report, independent from the project end-report which was drawn up by the project leader.

One or their findings addresses the lack of structured documentation on deviations from the standard workflow or migration-information. This was not in line with good practice in digital preservation and did not support proper provenance information. It also lead to time consuming actions in the analyses, for instance in the preparation of the migration from DIAS to DM and in several cases we needed to rely on personal memories of employees.

Therefore they recommended better and more extensive documentation of processes and decision making regarding digital collections designated for long term storage. They underlined the importance of a proper indication of the status of this documentation, proper version management and systematic documentation of findings is necessary for proper monitoring and checking afterwards.

Still to be done

Although all digital objects are now safely stored in the new DM, there still is work to do. One of the main activities will be to add the event metadata to each record in the Metadata store, based on the information in the Administration database. A design for the relevant Mets fields is underway. Another project will be to add the updated bibliographic metadata to the digital objects where there was no information in the KB-MDO (which is registered in the status in the Admin Database). This will be a manual action.

7. Conclusion

The migration of nearly 5 million objects from one almost obsolete storage and preservation environment to the next, future proof DM has proven successful. The project team kept to time and budget and lost almost no content. During the project we encountered several issues but were able to tackle most of them. The few post-project issues are currently being taken up.

There are issues to be considered. We only performed a relatively simple system- and data migration, no data conversion or file format conversion was involved. It safe to assume this would make a migration in the future more complicated, costly and time-consuming. Also, the project would have been very much larger had we wanted to also move the almost 6 million Elsevier articles.
Finally, the project taught us an important lesson to prevent problems in the future: better documentation of events in metadata.