http://conference.ifla.org/ifla78 Date submitted: 12 July 2012

WORLD LIBRARY
AND INFORMATION
CONGRESS:
78TH IFLA GENERAL
CONFERENCE
AND ASSEMBLY

IFLA
2012 Helsinki

LIBRARIES NOW!
INSPIRING
SURPRISING
EMPOWERING

Moving to new digital storage: migrating and reloading collections

Tanja de Boer and **Matthijs van Otegem**
Koninklijke Bibliotheek, the National Library of the Netherlands
The Hague, Netherlands

Session: **102 — *Storage and repositories: new preservation and access strategies* — Preservation and Conservation Programme (PAC)**

Abstract:

Ten years ago the KB e-Depot became operational; the first long-term archive for international scientific publications, worldwide. In these past 10 years, almost 18 million publications, mostly e-journal articles in PDF format, have been ingested, stored and made accessible on site. Our current DIAS system is now almost at the end of its natural life and a new digital storage system is under development, Digitaal Magazijn or short DM. KB is moving collections from DIAS to DM and its new storage infrastructure. Migration has started this spring and will finish in the fall of this year. For the KB this is the second large scale preservation action after a media migration in 2006.

Two years ago KB decided on a new strategy for collection care, compliant with the Open Archival Information System, an ISO standard for sustainable archiving. This strategy is based on the value attributed to a group of objects. We developed this method for both physical and digital collections and have presented the model for discussion at international conferences. This paper describes how this strategy helped us make and evaluate decisions about preservation actions, specifically the current migration of our digital collections.

Migration on such a large scale is expensive and time-consuming, and creates a moment for every organisation to reconsider processes and policies. Preservation policies need to be evaluated against the economic situation, technical innovations, and in our case the newly developed policy for collection care.

A large proportion of the collection in the e-Depot is published by Elsevier, an international publishing firm based in the Netherlands. Elsevier has offered to re-supply us with all their e-articles from the past ten years in a new and better format. By mutual agreement, we will reload most of the Elsevier-content instead of migrating it. Exceptions will be made for content that Elsevier can no longer supply (removed, discontinued and transferred titles). In addition, a selection of the original publications will be kept for historical purposes, to allow research into other aspects of these documents than the content itself.

A thorough test of a random selection of publications in our current e-Depot, compared to an identical renewed set as supplied by Elsevier, will give us the necessary information about the format of the objects, their technical content and their intellectual content.

As to the question whether reloading is a valid digital preservation strategy in general, KB does not have the answer presently. But we do wish to discus this issue with our peers, here at IFLA and on other occasions.

1. Introduction and context: New digital storage

KB's e-Depot has been operational for almost 10 years; the first e-journals were ingested in 2003. Its infrastructure is now outdated. We have been working since 2008 on a new, scalable digital repository to last us at least the next 10 years.

KB sets out to process and preserve multiple types of digital collections while the current environment is tailor-made for processing and managing e-journal articles. Our growing digital collection now mainly consists of scientific e-journal articles in PDF-format, but we also want to collect and store e-books, e-journals and other forms of born digital publications.

KB needs to upscale its processing and storage environment for processing at least ten times as many digital items in a limited time frame as it does currently; processing digital items that will be much larger than they are currently and storing and managing at least twenty times as many Terabytes than it does currently. The new DM is scalable so that it allows storing 'big data' in the near future. It enables the implementation of preservation tools, and tools for quality control. Finally a new IT-system should be modular and not linear as is DIAS.

In the future scientific publications will not be the largest digital collection. Most of our storage capacity will be taken up by our digitized collections; images of physical sources in our own collections or in the collections of other Dutch heritage institutions. Currently KB digitizes around 85.000 pages per day.

Functionality for identification, characterisation, format-conversion, in addition to newly developed preservation functionality is added to the system to ensure permanent access.

Software applications used in our version of DIAS (1.3) have reached their 'end-of-life'. Although all components are standard IBM products and are still supported, their current combination in DIAS is becoming vulnerable. The KB-IBM maintenance contract will expire per January 1 2013.[1]

The new digital storage facility is based on the OAIS model, as is our current e-Depot. The OAIS-framework is a conceptual model and is translated in to practical solutions in our library. The different components of the OAIS-model are identifiable in the KB-system.

New *Digitaal Magazijn* (DM) consists of three major modules.

- Workflow & Services: a set of services for a flexible and scalable ingest, access and preservation;
- Storage management & Infrastructure: a generic storage system and the technical infrastructure for ingest.
- Process data & Metadata: to manage administrative and technical metadata en process data and the development of management processes on these data;

The modular approach for our new digital storage was developed in close cooperation with our international peers[2] and extensively presented at iPres (International Digital Preservation Conference) in 2010.

[1] Hilde van Wijngaarden, Judith Rog, Peter Marijnen, *Building blocks for the new KB e-Depot*, iPres 2010

[2] Eight national libraries in Europe (Spain, Portugal, Switzerland, Germany, UK, Czech Republic and Norway)

2. Preservation strategy: Preservation levels

A significant characteristic of our strategy for permanent access is the introduction of different preservation levels. Preservation levels were designed in order to assess an appropriate and cost efficient method of preservation. Not every collection represents the same value to the library, not every collection is preserved for the same reasons and not every collection needs the same treatment to ensure permanent access.

Therefore preservation levels will be determined by the value attributed to parts of the collection.

So for example, for the digital collections, we use three levels of collection care.

1. Maximum: pro-active preservation

2. Medium: active preservation

3. Minimum: no active preservation, storage as-is

The first level, pro-active preservation, is characterised by a maximum effort to keep content, structure and functionality of digital publications for the future. Ingest is limited to certain file formats, and for these formats continuous access will be guaranteed. Quality control during the ingest process will be strict. We will secure future access in an authentic way.
We will maintain the content, structure and functionality in the future, and digital objects may be normalized. Normalization is a specific form of migration. It involves migration of digital records to a limited number of standard formats on their arrival at the KB.
On the second level, these principles will apply with limitations. Active preservation will be less time-consuming and more cost efficient. On this level future access in original file format will not be required. We will maintain the usability of the file as well as preserving it as submitted (bit-level preservation).

The third level is no more (or less) than bit-stream preservation. This level will, for example, apply to files converted for web access. We will make sure they are stored as delivered and are retrievable.

Principles for a preservation strategy

Since the publication of our current Strategic Plan 2010-2013 KB is fully focused on building and equipping the digital library. To support this strategy a new Collection Care Plan was formulated and agreed in 2010. It sets out a strategy for integrated, efficient and effective collection care for both digital and physical collections along the following principles:

- Integrated collection care for digital files and physical objects
- Classification of collections into larger unities
- Value assessment of collections
- Indicative risk assessment
- Differentiated levels of collection care
- Care redirected from the most valuable collections, to those where the biggest loss of value is expected

Differentiate according to value

It is impossible and unaffordable to apply the same level of care to all our physical kilometres and digital terabytes of collection. And that is not per se necessary. Not all collections are equally important, and not all materials are equally vulnerable. It must not be a matter of course that the most valuable collections receive the best and most care. The best care should go to those collection units where the greatest loss of value is expected.

To make these differences visible, rationalised selection is necessary. Our instrument of selection is value assessment. Research- and cultural values of different collections are identified, qualified and quantified according to a limited set of criteria. This is the starting point for prioritizing levels of conservation and preservation. The method of evaluating is identical for paper-based or digital collections.

Values

To structure differentiated collection care KB have divided digital and physical collections into lots, or collection units. We have identified 15 units in the digital collections (ranging from websites to licences), and 9 physical units.

These lots have been submitted to valuation by our collection specialists. We have defined the values applicable to our collections.[3]

We decided on four primary criteria:

- Informational value
- Aesthetic values
- Historic value
- Social value

Primary criteria are the basic values of a collection itself, without relation to other collections. A collection must meet at least one of these criteria to be accepted for conservation or preservation.

'Informational' value is about the content of collections as a source for research and about the objects themselves as carrier of information. 'Aesthetic' value is determined by the artistic value of a lot, visual appeal, design qualities or creative or technical excellence. It is obvious in many parts of our special collections. Also in the digital collections the look and feel, for instance in our web archive, is regarded as aesthetic value. 'Historical' value is based on the age of collections and/or on the way in which these collections are connected to important events in our national history. 'Social' value is identified when collections are f importance to one or more groups in Dutch society. Social values rise when this group plays a more important role.

Next to the primary criteria, there are four comparative or secondary criteria, determined by comparison with our other collections:

- Use
- Completeness
- Condition
- Provenance

The secondary criteria affect the weight of the primary criteria.

'Use' measures the actual use of a collection, be that lending out, reading, online access, digitize etc. 'Completeness' is the way in which a collection is complete, compared to other collections. A collection can be unique, very representative, or more complete than collections elsewhere. Completeness, uniqueness, rarity of a collection is complementary. 'Rareness' will, especially for digital collections rise as time progresses. 'Condition' measures the physical or digital state of collections, as well the state of the medium (book, newspaper, digital file) or its content. In physical as well as in digital collections the authenticity of the text plays a paramount role. 'Provenance', finally, refers to the person or organisation of origin. With digital collections, provenance or origin can give us guarantees about the authenticity of the collection, and gives authority to historic or informational value. Provenance functions as a quality-mark. Certain publishers of digital publication may be more trustworthy and deliver better files than others deliver.

[3] In doing so we have used the knowledge and methodology that is presented in the Australian publication *Significance*, published by the Heritage Collection Council in 2001. The digital version *Significance 2.0* was presented in 2009.

Assessing the relevance of primary and comparative criteria per category, results in a qualitative and quantitative statement about the value or significance of that collection. Expert-based valuation of KB-collections lies at the root of ascribing certain preservation- or conservation levels. For that, we have developed a simple but effective system of quantifying these values by ascribing points and multiplying outcomes. We are currently using this model and are optimistic that it will help us in quantifying the total collection value and the way in which the total value is distributed over the various categories.

Differentiate according to risk

The value assessment needs to be followed by a risk assessment. Based on expertise and experience we will set up a limited risk assessment to indicate where the biggest losses in established value are expected to occur in the future. For a lot that holds mostly informational values readability is of the essence, and therefore preservation will be aimed at, for instance, guaranteeing contrast. For a lot that has a high aesthetic value keeping the look and feel of objects is paramount.

Preservation- and conservation levels

The last step, after the value- and risk-assessments is defining a set of preservation and conservation levels, applicable to specific collections, which share specific values and a susceptibility to specific risks. The levels and the actions that go with them are aimed at preventing loss of value.

A value-based system for collection care will enable us, through matching values and risks to focus on the loss of value for groups of objects. By applying levels of preservation, we aim to give the appropriate care to keep our collection accessible for generations to come.

In other words, identification of values and relating risks to specific values will enable the KB to determine the necessary quality and quantity of care for all our collections. We will spend our resources in a more effective and well-argued manner.

3. Preservation actions: Migrating and reloading

These strategic principles are applied to the decision-making on how to preserve our collections in our new DM. Until now two preservation strategies were available: emulate or migrate. In this case, migration is needed not because the file format becomes obsolete but because the preservation system and software retires. Whereas in the former the object itself is changed intentionally, in the latter the object should remain unaltered. In both cases the primary values of the object should be preserved. The KB decided to adopt a third strategy: reloading as a preservation action for part of our collection.

In December 2011, the actual migration from old to new digital storage started in two steps. The first step was moving digital collections from the e-Depot to temporary storage facility. The second step started in June as collections were transferred from temporary storage to the new DM storage- and metadata environment. We have chosen this two-step approach to limit our dependency on the processing capacity of the old DIAS system. Around 6 million objects will be migrated in this way. However, almost two thirds of the current content of the e-Depot will nót be migrated, around 8 Tb. Ten million scientific articles from the international, Dutch-based publishing firm Elsevier, dating from the years 2003-2012 will remain in DIAS and will be reloaded in the new system straight from the publisher.

Why reload Elsevier?

Since 1995 Elsevier and the KB have cooperated in the field of digital preservation. Our first pilot electronic deposit system was tested together with Elsevier and their publications were the first to be ingested in our DIAS system. The collection consisted of both newborn digital publications and digitized copies of printed back files. Since 2010, Elsevier has been converting its scientific articles from the current Effect format to

Contrast format, which is fully XML. Elsevier has offered KB to re-supply 10 million articles, the largest part of our current Elsevier collection. This was a good opportunity and for a combination of reasons KB decided to accept Elsevier's offer.
In the early days of DIAS the process for ingesting Elsevier had to be developed. We created style sheets, advised the publisher on improving the metadata and successively in close cooperation over time we made it into a smooth, automated process. Though we still have all the original submission information packages (SIPs), migrating them will be quite a challenge as it is not a homogeneous set. For each different style sheet the migration workflow has to be adapted and it is likely that some style sheets will have to be rebuild to process the data from the first years (2003-2004).
A practical issue is the processing time of our DIAS system. It was developed to store data permanently and has lived up to this expectation for the past ten years, yet it was not developed to retrieve data by hundreds of thousands of items at a time. Migration of all the e-Depot content will take 17 months. The DIAS-system will not be operational after January 2013 and with that in mind we must finish migrating by then - a migration process can hardly be maintained stable and reliable over such a long period. It can be concluded that the migration of the Elsevier content will be time consuming, costly, and most of all, not without risk. Each migration bears the risk of damaging or even losing content and this risk would apply especially to the first content ingested in the DIAS system.
Finally, on a positive note the set to be reloaded has some advantages over the old set. Most importantly, the metadata is offered in a consistent and persistent metadata standard. This does not have any impact on the content itself, but it helps to reduce flaws in our ingest process and afterwards, to plan and perform preservation actions for the entire set in a consistent manner.
So for these reasons the KB accepted Elsevier's offer and by mutual agreement decided to reload the articles instead of migrating them.

Conditions to consider reloading

One consideration was to pin point the conditions under which reloading can be regarded as a true preservation action. The KB has invested considerably in digital preservation and has acted as an advocate putting this issue on the international agenda. We want to continue our efforts and remain a trustworthy partner, both in an international and a national context. By the action of just reloading content when migration is considered to be 'difficult' does not fit within our policy or the international OAIS-standard so we have pushed forward and adopted this for both DIAS and our new digital storage.
Our principles under which reloading can be used as a preservation strategy are basic and straightforward:

1. The primary values of the new set must be identical to the old set
2. The new set must be complete
3. There must be a positive business case: the risks for reloading should be lower than the risks for migrating
4. A sample set of the old content must be preserved

When it was decided to reload, lists were drawn up of journal titles that were transferred from Elsevier to another publisher or that were out of commerce and therefore not likely to be delivered. All these titles and their content will be migrated from DIAS.
In theory, all other titles still rest with Elsevier and are still on the market so should in theory be offered for reloading. After nearly twenty years of cooperation with Elsevier on digital preservation we are well acquainted and in our experience the Elsevier database has become very consistent. At the moment we hardly encounter any flaws in our daily ingest. We decided to put the statement that the two sets are identical to the test. To this purpose we used the technique known as acceptance sampling to detect the level of defects, if any. It was like looking for a needle in a haystack! in 10M articles we expect a level of 100 defects per million (DPM). The size of the sample to test this hypothesis would be unmanageable (nearly 300,000 items). Therefore we set the constraints on at

least for 99% identical, with a reliability of 95%, which resulted in a sample of 299 and no defects allowed. After checking these 299 articles and having detected no defects, it may well be the case (and is even expected) that the set is for far more than 99% identical, but the effort to prove it would be too high. Having set these boundaries, the outline of the business case compares reloading with migrating. In this kind of complex migration projects a defect rate of 1% would be quite low, but there are no similar situations to compare with as yet.
Finally, we have decided to preserve a sample of the old set. Having established that the set is complete and the content is identical, it may be the case that other aspects differ which might be of interest for future research. By preserving a sample, we can still facilitate for instance research into publisher's metadata policies in the early twenty-first century.

You may ask - What will happen with the 8 Tb Elsevier content that is not migrated? First, it will not be 'thrown away'. We will keep the DIAS-hardware (a PLASMON-server) and its content after 2013, but it will not remain an environment for long term preservation. The DIAS hardware just goes into 'cold' storage; the publications will, at least for the foreseeable future, be accessible when needed but otherwise kept cool and in the dark.

Conclusion

Yes, the KB has decided not to migrate but to reload part of its collection of e-publications. For the reasons mentioned above we think that in this case the decision is justified. But does this make reloading a valid digital preservation strategy in general?

Possibly: in this case it works out well, but evidently there are some very specific and local conditions. One can foresee that other publishers will follow Elsevier's example. They may well offer libraries their renewed and bettered publications. So the question is - Do we store them as well as their original versions? How may this reflect on future strategy? It is a question that needs to be raised and addressed.

We need to consider reloading as a preservation strategy for born-digital collections on the long term. Until now, the standard policy is to store it and keep the old set as well. In this way the amount of data to be preserved rapidly increases. Furthermore, as we move from platform to platform over time slight alterations may occur with each migration and if we want to preserve the former set the content has to be duplicated again. Within twenty years we would have multiple versions of the same set. As storage becomes less expensive every year, this would not necessarily be a problem (as long as it becomes cheaper fast enough compared to the growth of our collections). Still the question must be raised whether the extra cost outweighs the added value of having all these sets. Presently, the KB does not have the answers to all the questions, but we do have the wish to discuss the issues with our peers, here at IFLA and at other occasions.