An attempt at modeling differentiated storage for digitized collections: finding the balance between storage, costs and preservation of digitized publications[1]

Trudie Stoutjesdijk
Koninklijke Bibliotheek, Netherlands
Operations Department,
Collection Care Sub department
Trudie.Stoutjesdijk@kb.nl

ABSTRACT
The Koninklijke Bibliotheek (KB) digitizes the national collection of the Netherlands. Digitization leads to multiple versions of a publication: a digital access file, a digital master file, back-ups of the digital versions and the physical original publication. This in turn increases the need for storage capacity quickly. And raises questions like: Should all versions be stored? Do all the versions need to be preserved in order to ensure permanent access, and if so which ones should be preserved and how? Based on the collection care plan and the content strategy a differentiated storage policy is set up in order to establish a relation between the physical object and the digital counterpart(s). This method assigns value to different collection lots and is used to find out how to apply collection care in an efficient way.

Keywords
Storage policies; Collection Care; Permanent Access; Digitized collections

1. INTRODUCTION
As a national library the KB collects and maintains all publications that appear in the Netherlands, as well as a part of the international publications about the Netherlands. One of the large, labor-intensive challenges is to digitize all the books, periodicals and newspapers published in The Netherlands since 1470. Right now, nearly 10% (60 million pages) has been digitized. Digitization provides different versions of an object; therefore the amount of different versions of an object increases rapidly, as well as the storage costs. The most obvious approach, in order to reduce costs is to make sure that there are fewer copies of a publication. Questions are "Do we need to store all the versions of a publication? What representation of a publication is the object of preservation? Which ones do we want to remove, which one do we want to preserve? And which is most cost-effective method? " In order to answer these questions a close look at the current storage strategy is necessary. On the basis of the current collection care plan and archival storage system, a new storage model is proposed for digitized publications which distinguishes 5 different levels. Subsequently, there had been an investigation on potential costs savings and ways to use alternative solutions such as re-scan and conversion (or on-the–fly conversion) are possible.

2. KB MISSION
The mission of the Koninklijke Bibliotheek (KB), the National Library of The Netherlands, is to offer everyone everywhere access to all digital and printed publications that appear in the Netherlands. In addition, the KB fosters the establishment of a new (digital) information infrastructure. Close cooperation between the KB, academic and public libraries is essential to grant everyone in the Netherlands access to scientific information. In order to achieve this goal, a transition from physical to digital is necessary.

3. THE BALANCE BETWEEN COLLECTION AND MANAGEMENT OF DIGITIZED PUBLICATIONS.
Our Collection Development Program [1] underpins the goal to make the KB collections digital. The program makes it clear when to choose paper, and when to choose digital. It also explains the conditions of the strategy 'everything published in and about the Netherlands.' In 2003, KB's e-Depot became operational. It was designed to preserve the electronic publications of Dutch publishers, in agreement with the Dutch voluntary deposit guidelines. Archival Agreements were signed with Dutch and internationally operating publishers. Ten years later, the e-Depot system, DIAS, is at the end of its natural life and a new digital preservation system (DPS) is being developed, called Digitaal Magazijn. The new DPS is a scalable digital archive; it consists of three major modules: (Workflow & Services; Process data and Metadata and Archival Storage) which represent the OAIS model. In 2012, the KB migrated collections from DIAS to the new DPS. The next step is the development of new ingest workflows for the all the digitized collections and new born digital collections on the new DPS.

3.1 Collection development
The KB collects and preserves the printed and digital publications that are published in the Netherlands (e.g. the Netherlands Collection), has an important collection of special old manuscripts and early printed works, and a large number of digital databases and e-journals. Our collections are of great importance as source material for (academic) research, as background reading for university and professional education courses, and for everyone else who is interested in Dutch history, culture and society. The selection strategy with regard to digital content is described in the Collection Plan 2010-2013. The transition from printed publications to a digital format is key priority. The KB wants to digitize in the coming years all books, newspapers and periodicals which have been printed in the Netherlands since 1470. This is an endeavor that is beyond the capacity of the KB organization. We have therefore sought to cooperate, at first only with public parties, but later also with private parties. Public partners are the Dutch House of Representatives, university libraries – in particular those of Leiden and Amsterdam (University of Amsterdam) – and other

[1] With many thanks to Tanja de Boer, Irene Hasslinger, Barbara Sierman en Marcel Ras.

cultural heritage institutions. In this way all the parliamentary papers and more than 10,000 Dutch early printed books from the end of the eighteenth century have been digitized. Various national and foreign cultural heritage institutions (archives, libraries) contribute to filling the website Historical Newspapers with nine million pages of newspapers dating from the seventeenth century to 1995.The KB sought cooperation with private parties for the first time. ProQuest is scanning our early printed books till 1700 and Google is digitizing our copyright free books from 1700 to around 1870.

3.2 Collection Care

Storage is one of the main costs of Collection Care. . In order to guarantee permanent access to the digital cultural heritage of the Netherlands we need to store our collections as efficient as possible. Our Collection Development Plan is complimented is complimented by our Collection Care Plan [2] that sets out a strategy for integrated, efficient and effective collection care for both digital and physical collections based on the following principles [3]:

- Integrated collection care for digital and physical objects
- Classification of collections into larger unities
- Valuation of collections
- Risk identification
- Different levels of collection care
- Care redirected from the most valuable collections, to those where the highest loss of value is indicated

Table 1: Values

primary criteria	secondary criteria
informational value	Use
aesthetic value	Completeness
historical value	Condition
social value	Provenance

It is neither possible nor necessary to apply the same care to all the physical and digital collections. Simply because there are differences between collections and the care they need. Not all collections are equally important nor are they equally vulnerable. The best care should go to the collections for which he greatest loss of value is expected. In order to be able to value the collections KB have divided physical and digital collections into lots or categories. There are 25 different lots: 14 lots in the digitized collections; 9 in the physical collections. These lots have been submitted to valuation by the collection specialists based on the defined values in table 1.[2]

After value and risk-assessment a set of preservation levels for the lots can be defined. The preservation levels determines the actions that are aimed at preventing loss of value as well as focusing on the loss of value for group of objects. The goal is to give just enough care to maintain the ability to retrieve, view online and use digital material in the face of rapidly changing technology.

Table 2: Classification levels

Preservation level	1. Lowest	2.	3.	4.	5. Highest
Representation available?					
-Digital Master	No	No	Master light	Preservation master	Preservation master
- Access file	No	Yes	Yes	Yes	Yes
- Physical original	No	Yes	Yes	Yes	Yes
Preservation copy available?					
	No	No	Physical original	Preservation master	- Physical original - Preservation master
Replacement by representation desirable?					
	N/A	Access file	Access file and Master light	Access file	Access file

The identification of values and risks to specific values will make it possible to determine the specific nature and amount of care for all the collections. Resources will be spend in a more effective and objective manner[3]. At the moment we are working on the final value set of the lots. The emphasis in this paper is on the relationship between the physical original publication and the digitized counterpart(s). In anticipation of the outcome of the value proposition, this model for physical storage of digitized collections is mainly based on the (secondary value) condition of the physical collections. The digitized collections are not under threat because they are managed in-house; the specifications are drawn up by the KB and the file formats are known (TIFF and JPEG2000).

3.3 Finding the balance

There are many aspects that play role in efficient and sustainable storage of digitized publications. Digitization increases the amount of different versions of an object rapidly. Digitized publications will yield a physical, digital master and access version besides the back-ups (2 times). That raises the question what level of storage is needed for the different versions. The level of storage is also determined by the desired degree of sustainability for the various versions. And it is impossible to preserve all the versions at the highest preservation level and that will not be necessary. Finding the balance between these aspects is a real quest. Based on the collection care policy and the content strategy it is possible to establish a relation between the physical object and the digital counterpart(s) by assigning value to the different lots. This method makes it possible to apply collection care in an efficient way. There will be a distinct relation between the state of the physical object and the necessity of preservation imaging and sustainable storage of digital master files. A differentiated storage policy has been applied on the digitized collections; this is based on:

- The availability of digital contents for the customer
- The vulnerability of the physical resources
- The sustainability of digital storage

[2] Based on the Australian publication *Significance*, published by the Heritage Collection Council in 2001. The digital version *Significance 2.0* was presented in 2009.

[3] Tanja de Boer and Matthijs van Otegem. *Moving to new digital storage: migrating and reloading collections.* IFLA 2012.

3.4 Classification

Table 2 shows the classification of five levels, based on the values and the relationship between the different versions, the relationship between physical and digitized publications, the risks and the degree of effort that you want to apply to ensure permanent access of the collection(s). There is a distinction made between active and passive preservation; this only indicates which version is considered to be the master; that means the one that needs to be preserved for a long time. The concrete implementation of active and passive preservation effort needs to be completely based on the preservation policies.

Explanation of the different levels

Level 1: All these objects are available for use, the KB has no physical original and when usage drops, the subscription of the collection will be discontinued. This includes only licenses, there will be no conservation of objects in whatever form.

Level 2: This group is digitized to facilitate use. The main goal is the maintenance of the availability of the objects. The KB conserves the original objects passively: neatly stored on the shelf, whether or not compact stored in an air conditioned warehouse. The digital master need not be preserved. The digital derivative runs to the default backup and recovery procedure. This applies to all foreign titles in the Google project [4] (except if they still have value by particular provenance etc.).

Figure 1: Magazine Wendingen

Level 3: This group contains objects that represent multiple values. The physical object is in a quite good condition and can be digitized repeatedly. That is why a digital master does not need to have the high quality of a preservation image; neither should it be preserved in an active way. A digital master light[5] [5] would do. In this way it could save costs of production and storage. The digital master will be preserved in a passive manner; it runs along in the usual backup and recovery procedures. The original physical object will be actively preserved if necessary, in order to keep the value as an object available. This type of object is in both the special collections (large parts of the 18th century collections) and in the Metamorfoze period[6] (e.g. art books and cultural important magazines as Wendingen and De Stijl). Customers are working basically with the digital version; the paper version is available for specific research questions.

Level 4: This group contains objects with high information value. Full reproduction by digitization is usually possible. In some cases, however, the material could be so fragile and easily subject to deterioration that digitization could only be done once, and it will not be possible to maintain the physical original. In this case the aim is to create a preservation image at very high quality. This preservation master will be the retention copy instead of the original physical. This occurs in case of the Metamorfoze period where publications hardly represent value as an object.

No object is completely free from object value that is why the KB will not throw physical originals away. These objects will be conserved in a passive way: stored in an air conditioned warehouse. Customers will have to work with the digital version and only in exceptional cases access to the paper original is allowed.

Level 5: Only a small part of the collection is so precious, fragile or difficult to digitize that it can only be digitized once. It follows that the quality of the digital master should be as high as possible and maintenance is necessary, because there is no second chance to digitize. The physical object represents the primary values that might not be reflected in the digital master: historical, aesthetic and / or society: for example, a bookbinding of William the Silent, prints by H.N. Werkman (famous Dutch typographer, printer). Therefore the KB will preserve the original physical object actively. The customer can use the digital copy, but has, in many cases also access to the physical object.

Figure 2: Bookbinding William the Silent

3.5 In summary

The collections at the first and second level are exclusively for access. The first level exists of publications by subscription, the KB doesn't hold any objects only gives access to objects. The second level focuses on digitization for access only. None of them need active preservation and only level two needs passive preservation of both the original and the digital access file in order to keep them accessible.

A large difference can be observed between level 2 and 3, the context and reference collection on the one hand and the Netherlands Collection on the other.

Level 3 to 5 will require sustainable access at high level, either by active conservation of the physical object (level 3) or the digital object (Level 4) or both (level 5).

4. DIGITIZED COLLECTIONS AND STORAGE COSTS

In order to discover how to guarantee permanent access to the KB collections as efficient as possible one must have a clear understanding of the costs. There are cost models that cover the entire preservation lifecycle, these are all useful models[7], but there's still a strong development in the use of these models. One of the aspects of preservation is storage. In this model the focus will be on the storage model in use by the KB. For the calculation of

[4] http://www.kb.nl/sites/default/files/docs/contract-google-kb.pdf

[5] The Master light digitalization quality level is intended for digitalizing originals whereby color accuracy is slightly less significant. Examples include books, newspapers, magazines and hand-written material.

[6] http://www.metamorfoze.nl/english/home

[7] Keeping Research Data Safe (KRDS); Cost Model for Digital Preservation (CMDP); Digital Preservation for Libraries (DP4lib); Life Cycle Information for E-literature (LIFE3).

storage costs the KB uses a Total Cost of Ownership (TCO) for the entire storage infrastructure (including business and office storage costs[8]). Figure 1 shows the tiers and TCO of 2013 [4]; and the indicators for storage per TB. The storage costs per page are shown in table 3. The costs for digitization are based on a cost model and broken down by digital master files for permanent access and access files for current access. The following costs are distinguished: Capital costs, scanning costs and material costs. Based on this model and the production figures key performance indicators have been set (table 3).

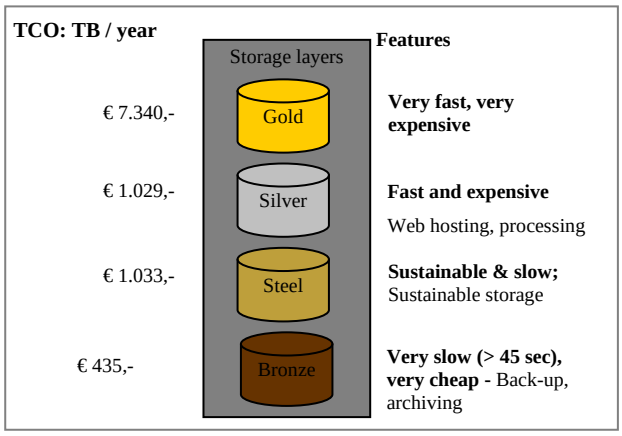

Figure 3: storage layers and costs

Cost Savings for storage of digitized publications

Currently the output of digitization process is a digital master and a digital access file. At the KB they are stored on different storage layers:
Silver: 1x digital access file
Steel: 1x digital master and in many cases 1x digital access file
Bronze: 2x back-up of tier steel (digital master and digital access file)

Table 3: Cost indicators

Type of publication	TB / page	Storage level / tier			Total costs / page	
Books		Bronze	Steel	Silver	Storage	Digitization
Master	0,00001	0,00435	0,01033		€ 0,01	€ 0,72
Access file	0,000001	0,000435		0,001029	€ 0,00	€ 0,56
Master & Access	0,000011	0,01914	0,01136	0,011319	€ 0,02	€ 1,28
Newspapers	TB/page	Bronze	Steel	Silver	Storage	Digitization
Master	0,00002	0,0087	0,02066		€ 0,02	€ 1,08
Access file	0,000006	0,00261		0,006174	€ 0,01	€ 0,93
Master & Access	0,000026	0,04524	0,02686	0,026754	€ 0,05	€ 2,01
Journals	TB/page	Bronze	Steel	Silver	Storage	Digitization
Master	0,000004	0,00174	0,00413		€ 0,00	€ 0,77
Access file	0,000001	0,000435		0,001029	€ 0,00	€ 0,61
Master & Access	0,000005	0,0087	0,00517	0,005145	€ 0,01	€ 1,38

[8] The components are partly based on the white paper "Four Principles for Reducing Total Cost of Ownership (2011 Hitachi).

At level 2 can yield cost savings because there will not be digital master files; this means that there are no production or storage costs for digital masters, only costs for digital access files. This could reduce the costs with 30 – 40%. At level 3 a digital master light will be created; a master light could require less image quality than a preservation master which could reduce the size of a digitized publication and lower costs of production and especially for storage. Digital master light criteria could be applied on objects of both the special collections (large parts of our 18th century collections) and in the Metamorfoze period (e.g. art books and magazines as Wendingen and De Stijl). But just now we do not have publications that are digitized conform the master light guidelines nor do we have cost indicators. As shown above, the application of the five level classification model reduce the storage costs of digitized publications.

4.1 Rescan and conversion to reduce storage costs

There can be several reasons for creating new digital master or digital access files. The access file no longer meets the requirements of the user, technologies offers new opportunities, possibly better and smaller digital masters or the original physical decay appears to be stronger than expected... Subsequently other additional methods to save costs were examined: rescanning and conversion.

Table 4: Rescan options

	Books		Newspapers		Journals	
Level 2: no master	storage	digitization	storage	digitization	storage	digitization
Master & access	€ 0,04	€ 1,28	€ 0,10	€ 2,01	€ 0,02	€ 1,38
Access	€ 0,00	€ 0,56	€ 0,01	€ 0,93	€ 0,00	€ 0,61
Savings € / page	4%	30%	10%	32%	9%	31%

Rescanning

Rescanning, i.e. re-digitization of (parts of) the collection, of an object is a way to get a new digital representation. Rescanning is only possible if the original is present and in good physical condition. Again, assuming a classification of five levels of retention, based on the relationship between physical and digital, the possibility and desirability for any rescanning be determined.
For objects of Level 4 and 5 (publications with informative value and object value) rescanning is undesirable and sometimes even impossible. Only the digital master can serve as a source for new derivatives. Rescanning is a costly affair, whether rescanning is done to create an access or digital master file. Therfore clear criteria should be drawn up to decide in which exceptional cases rescan schould be done. These criteria should reflect the wishes of the customer, the physical condition of the original and technological developments. For the manufacture of a digital access file, re-conversion can offer a solution, in particular for vulnerable physical collections. Rescan on the basis of the known data is not a suitable tool to use in the KB-storage strategy. The development of conversion and conversion on-the-fly can avoid rescanning.

Conversion and/or on the fly conversion

In this article conversion refers to the method to derive a new an access file from the digital master. A large part of the digitized publications is stored in JPEG2000. One of the guiding principles to use JPEG2000 was the ability to reduce the overall storage

requirements by creating smaller files. The digital JPEG2000 master can serve as source for the access files. There are several ways to deal with coversion, one can create access files in advance and store them in the same way as the current storage of access files takes place, or ceate an access file at the time a publication is requested, on-the-fly. Conversion on the fly will reduce the storage costs and will directly benefit those who want to use the KB collections online. But on-the-fly conversion has other objections, it is a system intensive activity that could create a bottleneck in the delivery to the end user [9]

For the collections that are classified at level 4 and 5, conversion and "conversion-on-the-fly" could be an appropriate and efficient method for storage and permanent access of the publications. In these cases there is no reason for rescanning. Conversion of digital objects seems to offer a considerable advantage of saving cost on production and storage of the digital access files on the expensive tier silver. There only needs to be one derivative to be generated at the time at a customer's request. There is little experience with conversion or on-the-fly conversion from digital master files to digital access files. This technique has not been applied yet. It is advisable to do research to determine whether conversion can be used for preservation and mobilization purposes. Therefore the research department is asked to investigate the applicability of this technique fo the digitized publications.

5. CONCLUSION

In this article a model is developed to reduce the storage costs of digitized publications at the KB. The model reflects a balance between collecting of publications at large scale, management of them for access and long-term, and costs. Finding this balance is important to keep permanent access of the KB collection affordable. Based on the current collection care plan and archival storage system, we proposed a new storage model for digitized publications with 5 distinct levels. By using this model it became clear which publications to preserve and how to preserve them. Transparency of the costs tells how expensive digitization and storage of publications are. It also gains a clear understanding of possible cost saving alternatives: reduce redundancy (do not store Digital access files on steel and silver, nor 4x on bronze), the creation of new digital master and/or digital access files by rescanning or conversion.

Rescanning is not feasible for publications that are in vulnerable state. Conversion might seem, from a cost efficiency point of view preferable to that of rescanning. Investigation of the conversion / on-the-fly conversion technique is necessary to gain insight into the benefits of this method. In particular with respect to applicability, performance and efficiency.

6. REFERENCES

[1] Koninklijke Bibliotheek, Collectieplan 2010-2013: Fysiek en digital integraal. Available from http://www.kb.nl/en/organization-and-policy/collection-development-programme-2010-2013 (2009) accessed 8 March 2013.

[2] Koninklijke Bibliotheek. Collectiebehoudsplan 2010-2013: Fysiek en digitaal integraal. http://www.kb.nl/organisatie-en-beleid/collectiebehoudsplan-2010-2013 (2009) accessed 15 March 2013.

[3] Boer, Tanja de, and Otegem, Matthijs van, Moving to new digital storage: migrating and reloading collections. In *78th IFLA General Conference and Assembly* (Helsinki, 2012). http://conference.ifla.org/past/ifla78/102-boer-en.pdf

[4] Hoeven, Jeffrey van der, and Zavaros, Rogier, March 20, 2013. KB Kennissessie : Expert meeting TCO opslag. http://intranet/kb-breed/kennissessies/eerdere-kennissessies/expert-meeting-tco-opslag-20-maart-2013; accessed April 4 2013.

[5] Dormolen, Hans, January 2012. Metamorfoze Guidelines: image Quality, version 1.0 of January 2012. http://www.metamorfoze.nl/sites/metamorfoze/files/bestanden/richtlijnen/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf

[9] Knijff, Johan van der, at jpeg2000wellcomelibrary.blogspot.nl/