LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, Printed in Germany

Practice, Organisation and Quality Control of Digitization Projects

by HENRIËTTE REERINK

INTRODUCTION

In the previous paper Dennis Schouten discussed the different aspects of preservation microfilming in projects carried out within the scope of Metamorfoze, the Dutch national preservation programme for library materials. I will do the same now for digitisation projects, which are executed within the framework of Metamorfoze. I will focus on the planning, workflow and implementation of Metamorfoze digitisation projects, on quality control and organizational aspects, and will relate these facets to preservation.

In 2001 Metamorfoze received a second four year subsidy from the Ministry of Education, Culture and Sciences to continue its national programme via its core-business of preservation microfilming. Part of the subsidy was earmarked for a new path, that of digitisation of collections that are already or will be preserved by microfilming also. With this new trajectory the ministry subscribes to the view of the hybrid method which is based on the technical connection between microfilming and scanning, or more generally, the coherence between preservation and digitisation projects: preservation via substitution microfilming and digitisation to increase access. It seems clear that combining the two processes offers considerable financial, organizational, practical and logistic advantages, in contrast to carrying out the two separately. Firstly, one can imagine that a combination saves time and that the knowledge of the structure of the collection and the metadata which is built up during preservation can be very useful in the process of digitisation. Secondly, a choice can be made for a high quality microfilm company which is also specialized in scanning and OCR. To keep both processes within one business, prevents problems such as that the microfilms to be digitised are not optimally suited for scanning. We have realized that there are missed opportunities when the two processes of microfilming and digitisation are carried out separately. Various activities have to be duplicated and because of the missed anticipation to digitisation, occurring errors need to be corrected or even ignored. Thirdly, for a filming and scanning business, building up knowledge of the collection while filming can also be very useful when scanning. Fourthly, when at the start of a project it is known that besides microfilming, also digitisation will be involved, one can plan for digitisation. Already in the microfilming phase the future digitisation has to be taken into account, which can have considerable consequences for specifications for the preservation microfilms - as we have learned from the *RLG Guidelines for Microfilming to support Digitization (*Dale, 2003) and the paper of Hans van Dormolen yesterday. And last but

not least, the physical condition of the collection will gain from a combination of preservation and digitisation. Handling is damaging to collections; in uniting the two processes, handling can be considerably reduced.

PRESERVATION & DIGITISATION

Digitisation projects are more complex with respect to planning and implementing than preservation microfilm projects. This is already evident in the beginning. With substitution microfilming a collection as a whole will be put on microfilm. The film ought to give a reliable and retraceable representation of all the original material, for that reason we speak of substitution microfilming. The microfilm exposures giving a near intact representation or measurable derivative of the source material. Since we cannot guarantee the long term durability of digital images and cannot yet measure the authenticity of the digital images in relation to the source material, digitisation is not yet a suitable preservation method. For these reasons substitution digitisation is not in question. In this case digitisation adds value only in so far it increases accessibility. Which in itself is valuable enough off course. This is one of the reasons that collections do not necessarily need to be digitised in their totality. The purpose of accessibility can also be served via an apt selection or maybe the collection is so specialist or obsolete in its sort that digitisation hardly serves any purpose at all. In considering digitising a collection some criteria are necessary. Which (parts of) collections do profit from digitisation and why?

METAMORFOSE CRITERIA FOR DIGITISATION

Metamorfoze uses a few criteria with respect to collections, which may be digitised. They are as follows:

- the collection must also be preserved via substitution microfilming
- importance of the collection
- frequency of use
- target group of users
- the physical state of the collection

If a collection is considered for digitisation and meets the above mentioned criteria to a sufficient degree, it is important to consider the keeping and maintenance of the digital

collection. In what kind of technical framework will the digital collection be stored and maintained?

MEMORY OF THE NETHERLANDS

For digitisation projects Metamorfoze works closely together with the Memory of the Netherlands programme. This is a cross-sectoral national digitisation programme to increase access to Dutch cultural heritage collections and is also subsidized by the Ministry of Education, Culture and Sciences. The Memory of the Netherlands has in the course of time developed a technical infrastructure for digital collections within the national library of the Netherlands, which provides a ready-made model for Internet accessibility and offers uniformity and structure. Metamorfoze could rely on this already developed and proven infrastructure.

INFRASTRUCTURE

By infrastructure I mean the way the scans and metadata are stored, retrieved and the way the specific collections are presented on the Internet. The infrastructure, which is offered by the Memory of the Netherlands, also sets certain quality standards for the scans delivered for presentation. The quality manager of the Memory of the Netherlands enforces these standards. He is the technical advisor and guard of the programme and he has developed digitisation standards primarily for the readability on the computer screen of the specific material offered. For the Memory of the Netherlands print quality or preservation quality, which would be essential if scans were made out of a preservation perspective, are of secondary importance. The main aim of the Memory of the Netherlands is broad access, long term preservation of the digital material is not yet an issue, nor is the measurable authenticity of the digital image in relation to the original material. The scans and metadata will be preserved and maintained for the near future in cooperation with the Koninklijke Bibliotheek, the National Library of the Netherlands. Before starting digital projects, these aspects relating to the infrastructure need to be thought over. Digitisation is by no means mere scanning of the original material. Since Internet access in itself is so easy, one often underestimates the efforts, which are required to present a collection in an organized way on the computer screen.

ADVANTAGES & DISADVANTAGES

In the case of the cooperation between Metamorfoze and the Memory of the Netherlands we noticed over time that adaptations on either side were required. The emphasis of the Memory of the Netherlands lay on the presentation of images with separate metadata for each image. Metamorfoze, as the national programme of preservation of library materials, deals more with text than image digitisation. The text material, moreover, often has only one single description for a cluster or compilation of documents or pages. These different angles of the two programmes needed to be attuned to each other. Also the OCR requirements of Metamorfoze were, by the nature of the programme, more specific than needed in first instance for the Memory of the Netherlands. This stimulated the Memory of the Netherlands to develop new tools and possibilities for this text-related feature. There appear to be many advantages and some disadvantages in working with an infrastructure which already has been laid out. I would like to repeat how basic it is with any digital project, to think these matters over before actually starting the project, and actually to reserve money for it.

PROJECT ORGANISATION

In Metamorfoze, with preservation as well as digitisation projects, the project falls under the responsibility of a temporary project leader from the institution which keeps the collection.

Besides the projectleader, a participating institution also supplies a cataloguer to the project (for the metadata) as well as someone with understanding of electronic data and how to retrieve them. For digitisation projects Metamorfoze makes use of the quality manager digitisation and the programmers of the Memory of the Netherlands. The participating institution remains responsible for the execution of the project. The project coordinator of Metamorfoze acts as advisor, controls the time-span of the project and represents Metamorfoze. He shares his expertise and experience with the project leader and the scanning business and sees to it that the project is carried out according to Metamorfoze standards and that it will fit into the infrastructure of the Memory of the Netherlands. He deliberates with the quality manager and the programmers of the Memory of the Netherlands. When the project is finished, he decides if it is completed properly. After the digitisation project is concluded, the digitised material will be kept, maintained and back upped by the Information Technology Department of the Koninklijke Bibliotheek. Be sure to always make agreements on this point before starting the project. The responsibility for the digitised material should be handed over from the project leader to a system manager of some sort. With Metamorfoze digitisation projects,

the Memory of the Netherlands, and the Koninklijke Bibliotheek guarantee the accessibility of the digitised material for the coming five years.

METAMORFOSE DIGITISATION PROJECTS

We do have four projects which involve manuscript collections: a collection of theatre songs from the first half of the 20th century, a collection of unpublished musical scores of the late 19th, begin 20th century Dutch composer Alphons Diepenbrock, an album amicorum of the Dutch 19th century woman author Truitje Bosboom-Toussaint and an archival collection of a Dutch 19th century regional poet who wrote in the French language. Of the printed collections we granted subsidy to a collection of children's books and a collection of anonymous brochures and pamphlets that were distributed during the Second World War. Scanning from microfilm and OCR is part of these digitising projects.

Also, we are working on a newspaper project. We will digitise a few national newspapers with diverse political and religious views from the period 1910-1930 from preservation film and apply OCR conversion. A combined project for periodicals is the fashion magazine *De Gracieuse*. *De Gracieuse* started in 1863 and ended in 1936. The illustrated periodical gives a wonderful insight into the fashion of past times in the Netherlands. The complete set of periodicals has been stored and microfilmed according to Metamorfoze standards. The colour of the lithographed plates however, was lost on the microfilm, since the project concerned greyscale, not colour filming. For digitisation we chose to scan from negative copies of the archive masters. The colour plates were scanned directly from the original magazine. These colour scans replaced the respective black and white scans made from the microfilm.

OUTLINE OF DIGITISATION PROJECTS

The activities involved in a Metamorfoze digitisation project are divided into five phases:

1. preliminary phase
2. metadata
3. digitisation
4. presentation
5. final report and evaluation

1. Preliminary phase

The steps of the preliminary phase are as follows:

- selection
- material analysis (types, numbers, sizes)
- copyright research
- request for a quotation
- judging offers and test results
- project proposal, estimate and planning

I will concentrate on 'the drawing up of a request for a quotation'. A request for a quotation covers many aspects:

- one company
- test: technical specifications (quality demands), filenames, (OCR)
- storage costs
- planning
- back-up
- treatment of the collection & insurance

The test is the most important aspect. It is essential for both parties, the institution holding the collection and the scanning business, that a scan test is made in the preliminary phase. The test should outline technical specifications as well as specifications for the filenames of the scans. The project leader in cooperation draws up technical specifications with the quality manager of the Memory of the Netherlands. Those specifications are adapted to the collection concerned and involve resolution, dynamic range, colour scale, file format and derivatives. If it is possible to scan from microfilm, technical specifications should also be drawn up for post scan image processing, such as removal of the microfilming targets, cropping, despeckling, deskewing and tone adjustment. If only a portion of all the preserved source material is selected, some scans must also be deleted. The contractor can do this upon detailed instructions. And this may be cost-effective with respect to the giving of filenames. Or the cataloguer can deselect the surplus scans later with the original material to hand. It is important to instruct the scanning company carefully about the filenames to be given to the scans. Every scan needs to be distinguished in the filename by a unique characteristic, which corresponds to the metadata. Besides, some scans are interrelated and this interrelation must be visible on the Internet. The mutual relation between these scans should also be expressed in the filenames. The assigning of filenames to the images must be included in the scan test. They must have a link to the metadata. For the test a small number of representative documents are selected or a microfilm with

exposures of difficult and diverse text material of the collection, depending on whether it is scanned from the original material or from microfilm. Be sure to illustrate the different type of filenames with an excess of examples. Experience has taught that it is always better to give too many examples than too little.

Some collections might profit from additional OCR. If OCR is an option, technical specifications and specifications for the filenames should also be drawn up for the OCR-ed files and should be included in the test for the contractor. A connection should be established between the OCR-ed files and the scans; in Metamorfoze projects we do this via the metadata.

2. Metadata

Phase two deals with the metadata:

- metadata research
- creating / complementing / correcting metadata
- writing conversion scripts for metadata
- selecting dump from database
- metadata control after conversion
- correcting conversion errors

I will focus here only at the metadata research. Investigating the structure of the metadata and the way they are (electronically) stored, is essential when analysing the material to be digitised in the preliminary phase. Somehow a connection needs to be established between the description of the document to be scanned and the actual scan itself when a document is presented on the Internet, because we want to be able to identify the scanned object when digitally presented. This connection can be made in different ways. With Metamorfoze projects we follow the scheme of the Memory of the Netherlands, which means that the connection lies in the filename of the scan. The filename often contains the shelf number of the document or any other unique characteristic by which this document is described. One can also give the scans a sequential numbering. Then a correspondence needs to be made in a database between these numbers and the actual metadata of the digitised material. Since every scan is different, unique identifiers of every scan are essential in this process.

3. Digitisation

Phase three consists of the actual digitisation, which, as you see, is just a mere part of the total project. In case it has been decided to digitise from microfilm, intermediaries of the microfilms have to be made. As outlined by Hans van Dormolen, Metamorfoze works with three generations of microfilms:

- a first generation archive master film of negative polarity (1N)
- a second generation duplicate or print master which mostly is of a positive polarity (2P)
- a third generation positive use or service copy (3P).

Because of the long-term preservation principle, the archive master is never used as an intermediary. We usually make negative copies of this first generation archive master for scanning from microfilm to have the least reduction in image quality. Duplicating negative archive masters can be expensive if it concerns many films. As yet we have not found a completely satisfying solution to this problem of cost-ineffectiveness. We are aware of this problem and working on it. It seems that greyscale films also give good scan results if scanned from positive instead of negative films. In that case, the positive duplicate master might be used as the intermediary for the scanning process. With a thorough preparation of the project the actual digitisation of the collection is the least difficult part of all the activities involved. If, however, the preparation is not thorough enough, quality controls can be very burdensome.

It is very important to control the results of the scanning business during the digitisation process and not only at the end. If the contractor is not scanning exactly according to the agreements in the quotation and the test, errors in the execution of the technical specifications or in the giving of filenames of the scans or the OCR can still be adjusted in an early stage. This quality control is essential and is in first instance done by the quality manager. For his quality controls he developed a control sheet, which you will find enclosed in your congress folder. His visits to the contractors are very effective, and can be seen overall as a long term investment. Contractors learn from our quality demands, wherefrom we in our turn profit in following projects executed by them for our national programmes.

4. Presentation: development site

Phase four of a digitisation project consists of providing the Internet accessibility and presentation:

- specifications for the site
- writing and editing introductory texts
- technical development of the site
- making test plan
- testing site
- adapting site following test results
- english translation of the texts

6. Evaluation

The last phase of the project consists of a written evaluation and a financial end report.

COSTS

A global impression of the relative costs involved in digitisation and microfilm projects, which are executed separately, and in combination can be made. Please note that these are very rough indications and that the actual costs off course depend of the specific projects.

Costs digitisation projects:

Digitisation	30%
Metadata	30%
Management	30%
Overhead	10%

Costs preservation microfilming:

Microfilming	50%
Organizing the collection	20%
Management	20%
Overhead	10%

Costs of combination:

Microfilming and digitisation	40%
Organizing the collection (incl. metadata)	25%
Management	25%
Overhead	10%

RESEARCH

In this paper I focused on several important issues, which you come across when combining microfilming and digitisation projects. Many issues are still in an indefinite stage, they have not crystallized yet in standards or satisfying solutions. More combined

projects need to be done and more specific research in each of these issues is necessary. Metamorfoze has planned a desk research for the coming half year about the state of the art of combining microfilming and digitisation. The research will include a survey of combined projects carried out worldwide. It will examine possibilities and impossibilities of the COM-method, hybrid camera's and standards for scanning from microfilm now and in the near future. Attention will be given to themes such as scan targets to measure the degree of deviation and manipulation of the scanned source material in relation to the COM-method. We want to find out if hybrid camera's are being used and if their results are satisfying for preservation and digitisation goals. And we will analyse quality differences resulting from scanning from either positive or negative greyscale film copies. The afore mentioned *RLG Guidelines* will also be taken up in this research. My experience is that as you try to realize a combination of the two processes, more questions arise than answers are given. I hope that this presentation has contributed to raising more questions. And even more I hope that our combined conference and workshop will lead to some answers.

REFERENCES

Dale, Robin L. *RLG Guidelines for Microfilming to Support Digitization*. RLG, 2003. http://www.rlg.org/preserv/microsuppl.pdf

WEB SITES REFERRED TO IN THE TEXT

Koninklijke Bibliotheek. http://www.kb.nl/

The Memory of the Netherlands. http://www.geheugenvannederland.nl/

Metamorfoze. http://www.kb.nl/coop/metamorfoze/home.html