Using Microfilm for Long-Term Preservation of Digitally Annotated Archival Content

Alexander Stenzer [(1)], Burkhard Freitag [(1)]

[(1)] University of Passau
Chair of Information Management, 94030, Passau, Germany
EMail: {alexander.stenzer, burkhard.freitag}@uni-passau.de

ABSTRACT

Digitized versions of valuable artefacts and archival items, like plans or pictures, can be stored and enriched with metadata using digital archiving systems. Long-term preservation is of course a major issue that has to be addressed. In this paper, we show how the Scan First – Film Second method combined with a hybrid approach, i.e., microfilm plus digital data, is be applied to the MonArch Digital Archiving System, an existing operational system for the digital preservation of cultural heritage. Following a hybrid approach, digitized versions of artefacts and archival items enriched with metadata are kept in safe digital storage and are additionally printed on microfilm for long-term preservation. The Scan First – Film Second approach ensures that all metadata attached to the digitized artefacts is also preserved onto microfilm. We present and discuss the entire workflow from physical archiving over digital data to microfilm and back to restored digital data including the metadata. The reconstructed digital items are then automatically re-inserted into the archiving system.

Keywords: long-term preservation, digital data, metadata, hybrid, microfilm

INTRODUCTION

Along with digitisation new challenges in the field of long-term preservation of digitally created objects arise. The main question is how the digital data can be read and used in the future, e.g., in 100 years.

This paper presents an information migration approach onto microfilm which can be used for digital image repositories. Using microfilm for long-term preservation is a well known and established method. Black and white microfilm is said to last about 500 years if it is stored appropriately [1]. For digital file formats no reliable experiments exist telling how long their durability really extends. The main challenge faced during developing the approach presented in this paper was the fact that the transfer onto microfilm causes a certain loss of functionality [5] and the loss of all metadata of digital images. This is why a workflow which first makes a microfilm from the archival objects and then scans the microfilm to create a digital representation is not suitable to preserve digital images with attached metadata. How important metadata is for cultural heritage resources is e.g. shown in [2].

In the first section, the end-to-end workflow is presented in detail. The second section is devoted to the MonArch use case which includes a prototypical implementation of the entire workflow. The paper ends with a short summary.

WORKFLOW SCAN FIRST – FILM SECOND

The starting point is a conventional archive which may contain architectural drawings, photographs, text documents, and other physical artifacts. These archival elements belonging to our cultural heritage are digitized for further operations. Digitization allows for an easy access for large groups of interested people like scientist, architects or the general public while avoiding the archivals to be worn out by use. However, the advantages of digital representations are not in the focus of this paper.

In the first step of the workflow (see Figure 1), the archival elements are digitally scanned or photographed and stored as TIFF images in a repository, e.g. the MonArch Digital Archiving System [3].

During this step, technical metadata such as date of digitization or image resolution is created automatically as usual. In contrast, "semantical" metadata such as the name of the author of a text or the original creation date of a drawing can be added manually. Further annotations, describing, for instance, style or epoch can be added using hierarchically organized keywords predefined by specific ontologies. Often, also the geographical or structural position of an artefact is considered relevant. Thus spatial references model another kind of metadata. The stored metadata can be used by different kinds of retrieval mechanisms, like keyword search or spatial queries. In a next step of the preservation workflow, the current state of the archiving system is prepared for long-term preservation onto microfilm. To this end, the archival items have to be exported from the archiving system. The images are exported as TIFF-files whereas the metadata of every image is exported as an XML-file using Adobe's XMP format. Metadata which is relevant for more than one document or even all documents in the archive is exported separately. The transfer of the digital data to microfilm is the main task of the next step. Recall that in our use case digital data consists of both image information and meta information. Consequently, these different parts have to be transferred and every part of the digital data has to be represented on the microfilm in a suitable way. For our purposes, grey scale microfilm is preferred due to the following reasons: First, grey scale microfilm has low manufacturing costs. Second, the experience in handling the microfilm is way more advanced and third, its durability is assumed to be better than that of colour microfilm. Therefore a colour separation, e.g. resulting in three RGB frames, is necessary for the image information [4]. The metadata information is transferred to an extra frame on the microfilm. In [6] different kinds of representing metadata on microfilm are proposed, e.g., using a two-dimensional barcode. However, when attempting to restore data from the microfilm information on how to decode the two-dimensional barcode is needed. The same problem occurs with file formats. If a digital file has to be restored from a storage medium the information how to interpret the file format is necessary. The methods proposed in [6] have one drawback in common: they are not human readable. To overcome this issue we decided to represent the metadata on microfilm in such a way that it is human readable. In the presented case, the metadata is coded in the XMP format and stored as XML on the microfilm frame. We believe that relying on XML in combination with XMP that it should be possible to restore the metadata even without knowing how to decode the metadata format. To do so, the metadata is represented as an image containing text on a separate frame. Both, metadata concerning a specific image and metadata concerning the whole archive, are stored on separate frames the same way.

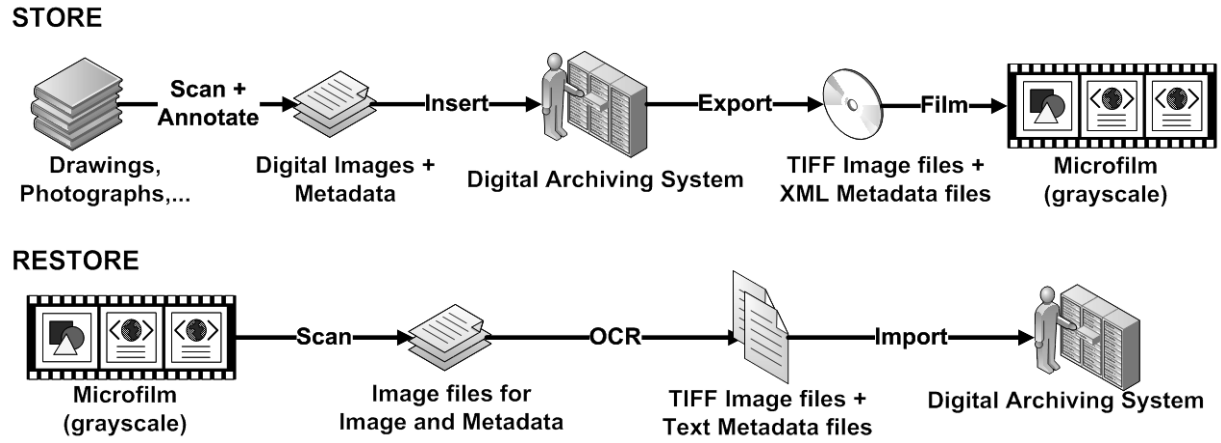

Figure 1: Entire workflow for long-term preservation onto microfilm

To restore the digital data from microfilm the entire microfilm has to be scanned. For each frame one image is created. The images can then be sorted such that for each image the grey scale image frames described above, several metadata frames - depending on the extent of the metadata - and several frames for metadata which are relevant for more than one document in the archive or for the whole archive are extracted. Secondly, the three grey scale image frames are combined to one colour frame and stored as a TIFF-image file. Next, OCR is used on the metadata frames of each image to reconstruct the textual information. The same OCR method is used to restore the metadata concerning more than one document or the whole archive. Validating the metadata for each image and the extracted text against the XMP

schema is the next step. In the final step the image file combined with its metadata is imported into the digital archiving system. As a result, the complete archival content is restored from the microfilm.

A disadvantage of our proposed workflow could be that the quality of the restored results fully depends on the quality of the OCR-technique and its error rate. Our proposed workflow does not fit for scenarios where digital archiving systems are subject to permanent changes and where different versions should be stored on microfilm on a daily or weekly basis as it would be too time-consuming and too expensive. The workflow fits best for digitization projects where a huge amount of analogue pictures or drawings has to be scanned, the digital representation has to be inserted into a digital archiving system annotated with valuable metadata and where a state worth to be preserved is reached after a certain amount of time.

USE CASE: MONARCH

The MonArch Digital Archiving System is a system which indexes digitized documents based on their structural and spatial position as well as on semantical properties organized as an ontology. As a consequence of its features the MonArch Digital Archiving System is well suitable for an illustrating use case serving as a basis to prototypically implement the long-term preservation workflow. In the MonArch use case, the content mainly consists of large drawings, historical pictures of buildings and text documents. The digital representation of the content is stored in a MonArch archive and the semantical meaning of the content is described by specific ontologies. As described above, there are metadata attached to single documents as well as metadata applicable to the entire repository.

The functionality needed for the first step of the previously described workflow is part of the basic functionality of the MonArch Digital Archiving System. The export functionality which has been added takes all TIFF-images stored in the MonArch archive and saves them into a TAR-file. For each image a separate XML-file is created which contains the metadata as XMP. The separate metadata XML-file is also stored in the TAR-file, the used ontologies are exported as separate XML-files and added to the TAR-file. To ensure that the image file and the corresponding metadata file can be identified within the TAR-file, both file names are replaced by the same generated unique identifier. The TAR-file can be handled by an expert in the field of microfilming. It contains all necessary information to create a microfilm which contains the image information and the annotated metadata.

CONCLUSION

In this paper a workflow for long-term preservation using microfilm has been proposed that is suitable for documents, in particular digital images, which are annotated with various kinds of metadata. First experiments have shown that a lossless restoration of the original digital data can be achieved.

REFERENCES

[1] M. Castillo. Preservation of knowledge, part 1: Paper and microfilm. *American Journal of Neuroradiology*, 30(9):1627, 2009.

[2] Milena Dobreva and Nikola Ikonomov. The role of metadata in the longevity of cultural heritage resources. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 69–76, Stroudsburg, PA, USA, 2009.

[3] Burkhard Freitag and Christoph Schlieder. Monarch - digital archives for monumental buildings. *Künstliche Intelligenz (KI)*, (4):30–35, 2009.

[4] Joseph E. LaBarca. Image storage and permanence considerations in the long-term preservation of photographic images – update 2010. *Journal of Physics: Preservation and Conservation Issues in Digital Printing and Digital Photography*, Conference Series 231, 2010.

[5] K.H. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary. The state of the art and practice in digital preservation. *Journal of Research-National Institute of Standards and Technology*, 107(1):93–106, 2002.

[6] Steffen W. Schilke and Andreas Rauber. Long-term archiving of digital data on microfilm. *Int. J. Electronic Governance*, Vol. 3, No. 3:237–253, 2010.