The Digitization of the Slide Collection from the Saint Catherine Library Conservation Project

Athanasios Velios and Nicholas Pickwoad; Camberwell College of Arts/The University of the Arts, London; UK

Abstract

Saint Catherine's Monastery in Sinai holds one of the most important collections of Byzantine manuscripts in the world. Camberwell College of Arts has completed a detailed condition assessment of the manuscripts, and has collected photographs of the bindings on colour transparency (slide) film. In this paper, we explain why we chose film photography rather than digital, we describe the methodology of digitizing the slides and we explain how low-cost equipment can be used to produce digital images without compromising quality. We use the JPEG2000 format with a combination of Dublin Core for descriptive metadata and the DIG35 standard for technical metadata. We conclude with a short discussion of the limitations of our methodology and the resources available.

Introduction

Saint Catherine's Monastery in Sinai holds one of the most important collections of Byzantine manuscripts in the world. It is important not only for the palaeographic value of the manuscripts but also because of the large numbers of original bindings preserved in the extremely dry and remote conditions of the desert. Camberwell College of Arts has completed a detailed assessment of the condition of the manuscripts and the information collected is kept in a database which is an invaluable tool for planning conservation work at the library. Part of the condition assessment was to record the bindings on slide film.

The slides are currently stored in filing cabinets using plastic sheets with pockets for each individual slide. A large number of them have been annotated. The slide title and film roll number is marked on the frames of the slides. In order to be able to combine the data from the database with the information captured in the images, funding from the Headley Trust based in the U.K. was offered to enable the digitization of the slides and their storage alongside the database records.

This article describes the methodology used during the slide digitization and the various software and hardware tools. The article also discusses the decisions made to implement that methodology and some considerations for the technology chosen. We begin our description with the photographic material which was digitized.

Photographic records

Photographic recording of the manuscripts is an important part of the condition assessment. In this section we explain why we have chosen to use film photography as opposed to digital and we describe how the manuscripts were photographed during the condition assessment.

Slide film

As mentioned in the introduction, the condition assessment of the manuscripts started in 2001 but had been planned earlier. At that time digital photography had already been introduced in various fields and we faced the dilemma of using either digital or film cameras. On the one hand digital cameras were a promising technology with the obvious advantage of rapid results. On the other hand film cameras had been proven to be reliable and it was equipment that we were confident with. However the main reasons why we opted for film camera instead of digital are:

1. Equipment maintenance. The remote location of the monastery makes equipment support impossible. Therefore, we could not afford a possible malfunction of the digital camera. Furthermore, the desert dust penetrates almost every piece of equipment and digital cameras would be vulnerable to this. Film cameras, on the other hand, have guaranteed reliability and are simpler to clean than digital cameras.
2. Archiving considerations. At that time, few organizations could afford the expertise to create a future-proof digital archive. The risk of compromising the quality of our archive led us to choose the safe option of archiving slide film (Kodachrome 64) which again had proven stability and was a well-researched medium.
3. Consistency. The assessment was planned to last for about 5 years. Within this period of time, digital camera technology progressed significantly and if we had chosen the digital option, inevitably we would have been working with obsolete equipment soon after the beginning of the project. Slide photography ensured the consistency of our records.
4. Cost. Although we have spent a significant percentage of our budget on slide films, arguably this was the cheaper option. Digital high-resolution cameras in 2001 were far more expensive than they are now. Also, storing and backing up thousands of images at the monastery would have added an additional expense and if we had to use and support personal computers at the monastery during the photography, that would also have increased the expense further.

For the above reasons we chose to use slide film instead of digital photography. Table 1 shows a list of the specific equipment that we used. Of course, the current photographic equipment available and the developments in digital archiving might make us revisit our decision if the choice was to be made now.

Photographic record

A detailed description of how the condition assessment of the manuscripts was done has been given by Pickwoad [1]. Here we will only describe the photographic records.

Eight shots are captured from each manuscript. These are photographs of the exterior of the boards (covers), the spine, the foredge, the head (top), the tail (bottom) and the interiors of the

boards (inside the covers). Manuscripts in bad condition may have their boards missing, in which case there are only six photographs of the manuscript taken. Figure 1 shows drawings of the eight different shots. Alongside these standard photographs, details of interesting binding features or representative examples of damage to the manuscript are also taken. These depend on each individual book and it is up to the assessor to decide whether they are worth taking or not. Each photograph is logged on a special paper form. The log form includes the shelfmark of the manuscript, the description of the photograph (e.g. Left Board Interior), notes about the aperture and shutter speed of the camera and the film roll number (unique for each roll). We have used flash photography and the same type of slide film for all photographs, so it is not necessary to include this information for every shot as it remains unchanged. The log sheets are the initial record of both descriptive and technical metadata of the image.

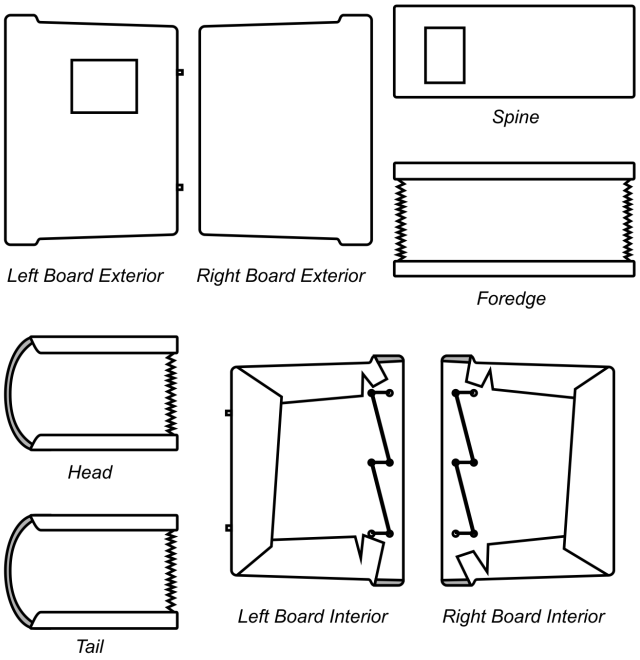

Figure 1. *Schematic representation of the eight standard photographs taken, during the assessment of the manuscript's condition.*

Colour management is achieved by photographing our colour chart in the first frame of each roll using the same setup. Any colour correction can then be performed by comparing the colour of each slide to the scale of the first frame of the corresponding roll. Having collected the images on site we then digitize them in our office in London. In the next section we describe the digitization technologies we use.

Image archiving technologies

There is a wide range of tools available for digitizing photographic records and slides in particular. These tools vary in quality and cost. In this section we will explain how we chose archiving technologies which, although economical, were at the same time aligned with current trends in digitization.

Scanner

When choosing a slide scanner, we faced two options: buying a) an expensive and fast scanner or b) a cheaper but slower scanner. As mentioned before, our slides are kept in plastic pockets within large file cabinets and they are annotated when removed from the cabinets in order to be scanned. Removing the slides from the cabinets one by one, annotating them and feeding them into a scanner is inevitably the slowest part of the digitization process. Therefore, having a truly fast scanner would not actually make the scanning process any faster, as the scanner would be idle while the user fed, annotated and removed slides. For this reason, we chose to use a slower and cheaper scanner. During the scan, the user is busy with preparing the next batch of slides and although scanning takes longer, time is not wasted.

The resolution with which we scan is 3000 dpi. Tests revealed that scanning in higher resolution dramatically increased the disk storage demands without significantly improving the detail that was captured off the slide. Figures 2 and 3 show the same detail of a slide in 3000 and 4000 dpi. Although the 4000 dpi image is larger the information captured is not any clearer than the 3000dpi image. By scanning in lower resolution the cost of hard disk storage dropped, without loss of information from the image.

Figure 2. *Detail of a photograph from a slide of manuscript Arabica 175 scanned in 3000dpi.*

Figure 3. *The same detail as in Figure 2 scanned in 4000dpi.*

Image format

We have chosen to use JPEG2000 for the final storage of our images. The main reason for this choice was the potential for lossless compression that JPEG2000 offers, which significantly reduced the need for storage space on our server without compromising the quality of our images. If we had used TIFF files instead of JPEG2000 we would probably have needed at least twice the currently required hard disk space. In addition, JPEG2000 offers other important capabilities, such as metadata storage within the image file [5] and the network-friendly tiling feature which allows partial loading of the file [6]. A low resolution compressed JPEG version of our images is stored temporarily alongside the JPEG2000 images, for quick retrieval with currently available tools as explained later in the article.

Metadata

The information from the condition assessment for each book kept in the project's database will ultimately evolve as a set of metadata which will describe the binding structure and condition of a manuscript. This metadata will be ideal for annotating the images. However at the moment this work has not been completed. Therefore, we decided that it is not advisable to keep detailed metadata about the manuscripts depicted in the images. Instead we keep general metadata about the images only. We use the widely accepted Dublin Core (DC) set of metadata. Most of the tags used for DC are adequate for describing the images (i.e. shot description) and their copyright. We map the DC as follows:

- Title: Manuscript shelfmark followed by a ".", followed by the number of the shot. The number of shot corresponds to one of the eight standard shots as described earlier. Images of details always start at number 09 and continue sequentially.
- Creator: Name and surname of the person who digitized the slide as opposed to the person who shot the initial photograph.
- Description: The setup corresponding to the number of the shot (e.g. Left Board Interior). When we have images of details, this field corresponds to the description of the shot as logged on site.

The rest of the fields are common for all images and mainly name the publishing body which is Camberwell College of Arts and naturally the Monastery which is the owner of the copyright. A sample XML file of our metadata can be downloaded from the Project's website or at the address:

www.arts.ac.uk/research/stcatherines/files/sample.xml

Having explained the use of Dublin Core for our content description, we will now focus on the technical metadata. A number of different metadata sets can be used for image description. These are either abstract, such as PREMIS [2] or more specific such as EXIF [3]. Large digital object collections benefit from abstract metadata sets, because such sets can describe a variety of digital objects. Our digital objects only include images and therefore we decided to use more specific metadata for their description. EXIF was on obvious option; however EXIF is strongly oriented to images produced by digital cameras whereas in our case, images are produced by a scanner. For this reason we chose to use a metadata standard proposed by the Digital Imaging Group (DIG), namely the DIG35: Metadata for Digital Images [4]. The advantages of this standard are summarized here:

1. Support of metadata for scanners. DIG35 contains a set of metadata tags which are specific for scanning equipment and include descriptions of the original medium (i.e. film). Although, in DIG35 the scanner metadata is not as detailed as the digital camera metadata, we found that it is difficult (if not impossible) to extract detailed technical metadata from the scanner while a scan is being performed. We discuss this later in this article.
2. Industry support. DIG is supported by many major companies which are active in both the fields of computing and photography. For this reason we believe that the DIG35 standard will be supported widely in the future.
3. Easy to use. The high quality of the documentation of the standard and the detailed description of the specification with multiple examples, make DIG35 an option which is quick and simple to implement.

Again an example of a metadata file that we have produced can be found on the Project's website, at the address mentioned earlier. In the next paragraph we describe how both JPEG2000 images and XML metadata files are stored.

Storage

Images are stored on the Project's server, as separate files on the hard disk. A relational database holds references to these files and their correspondence with the bibliographic manuscript records (i.e. which images correspond to a specific manuscript). Metadata is stored in two locations: a) in the database as long textual information and b) inside the JPEG2000 file using the XML box capability of JPEG2000. This means that any changes to the metadata demand re-inserting the XML box in the JPEG2000 image. However, we decided to keep the metadata inside the image as well as in the database, in the unfortunate event that the link between the files and the database records might disappear.

We follow standard backup routines on tapes for securing the data.

In the next section we will explain how the technologies described above have been combined to form a simple methodology for digitization.

Digitization methodology

Slides are digitized on a per manuscript basis. This is because they are physically stored in plastic sheets with pockets and each sheet corresponds to a manuscript. To avoid confusion we work with one sheet at a time. Slides are then annotated if necessary and loaded onto the scanner where they are scanned in TIFF format and stored temporarily on the local disk. An in-house utility then picks the temporary files, collects the metadata and produces an instruction file. This file stores necessary information for the conversion from TIFF to JPEG2000 and storage of the final files on the server. This transformation takes place overnight, as a separate automated job, because it takes too long to perform while scanning. Let us now describe these steps in more detail.

Scanning

The slide feeder accommodates up to 50 slides per job which is more than adequate for the number of slides that we shoot per manuscript (i.e. rarely over 15). The slides can only be fed landscape through to the feeder, which means that the resulting images need to be rotated when necessary. Our scanner has special functions for auto-focus and auto-exposure. Auto-focus takes a very short time and is critical to the quality of the final image. We

therefore perform auto-focus on each slide separately. However, auto-exposure demands a longer time to perform and almost doubles the overall length of scanning time. As mentioned before, our slides have been shot in consistent lighting conditions and there is therefore an insignificant difference in the exposure among them. In order to keep the scanning time short, we decided not to perform auto-exposure on each slide separately. Instead we only use the auto-exposure feature at the beginning of each scanning session and keep the same settings throughout.

Another reason why we did not insist on using individual settings for every slide is because, after contacting the manufacturer, we discovered that it is impossible to record any of the auto-exposure settings when scanning and technical metadata for the images was therefore impossible to collect.

We do not use any software filters, such as sharpness, on the image. The only manipulation performed is for colour correcting the blue hue of the scanned image which is due to the Kodachrome film used and has also been observed elsewhere (e.g. [7] or [8]). The blue hue is consistent and the colour correction is done by using the colour scale photographed on the first frame of each film roll. The settings are kept unchanged throughout the whole set.

We are using the sRGB [9] colour space as our images are mostly meant to be examined on screen and sRGB is often used in monitors. Moreover, our archive does not serve as colour reference but as reference for the binding structures, hence it is not our intention to control colour more accurately. Therefore, sRGB is an adequate colour space for our needs.

Our slides are kept in sequence according to the eight standard shots for each manuscript. When slides are missing, and hence the sequence is broken, they are replaced with blank slides which are discarded during the metadata creation. We have found that it is faster to scan blank slides to complete the sequence rather than to use software for rearranging the slides to match the sequence.

Having stored the images in a temporary folder as TIFF, metadata collection is our next step.

Metadata collection

In order to minimize the time needed for metadata inputting, we have developed a utility which assists the user in this task. The utility's functions are divided into four groups as described next.

Loading images

The utility has been designed to import images from a range of sources including a folder on the disk. Images are loaded as references to the original files on disk which are not altered at this stage.

Producing metadata

Because we are using a consistent process for scanning our slides, technical metadata is almost identical throughout the whole collection. Information about the date that the image was taken at the monastery, alongside the date that the resulting slide was scanned, is collected on-the-fly from our database records and by retrieving the current system date. Descriptive metadata is manually copied from the photography log sheets. The fact that there are eight standard shots per manuscript means that the descriptive metadata is identical for most of our collection and hence automatic metadata inputting can be used. Manual inputting is only demanded when extraordinary shots are taken out of the standard sequence (i.e. details of interesting features or damage). Otherwise the sequence of the eight slides defines automatically the description of the specific image. For example the first slide of a manuscript is always the Left Board Exterior shot, the second is always the Right Board Exterior and so on. This is why it is important to scan the slides in sequence. As mentioned before, when slides are missing, blank slides are scanned to complete the sequence and the resulting void images are discarded. This is faster than the user trying to resolve the correct sequence on screen and also less prone to errors. Other metadata information includes the rotation angle of the image in order to turn portrait images which have been scanned as landscape. The slide exposure and lens aperture alongside the film roll number are kept on the log sheets and are also recorded as part of the metadata. Finally, each digital file is given a unique identification number (UID) which is obtained by our utility after querying an external application (see list of tools) during the automatic overnight processing.

Exporting instructions

The metadata information is kept in a simple text file alongside the original TIFF file. The file only stores unique metadata for the specific image. The metadata which remains unchanged for the whole collection are not included here. Until this point all files are held locally (not on the server).

Processing files

The scanned TIFF images need to be converted to JPEG2000 following the instructions and including the metadata stored in the simple text files. We discovered that due to the huge size of our images, conversion takes too long to perform while scanning. Also, the increased load of our server during the day and the busy network would further delay moving the files to the server. For this reason the conversion takes place overnight, when both the network and our server are not busy, using an automated script. The script follows these steps:

1. gets a TIFF file and text instructions for it from the disk,
2. requests a UID from an external utility for the image,
3. produces an XML file with the Dublin Core and DIG35 metadata,
4. calls an external utility which converts the TIFF file to JPEG2000 and encodes the metadata as an XML box [5] in the JPEG2000 file,
5. produces a low resolution highly compressed JPEG file of the image, (we discuss why we produce this additional file in the next section),
6. copies the JPEG2000 and JPEG file from the local disk to the server over the network,
7. creates a reference in the database for the JPEG2000 image and stores the metadata file in a designated database table as text,
8. removes the file from the local disk and
9. continues until no images are left.

The metadata produced records the history of the digital file from its generation as a TIFF scanned file to the colour-processing, rotation and conversion to JPEG2000.

The methodology described above, allows us to digitize an average of 7 sets of slides (about 55 slides) per hour. Working at this rate, our part-time digitizing staff will be able to complete the

digitization of about 33000 slides of the collection in about a year. Although this methodology works well with the resources currently available, in the next section we discuss some disadvantages of our approach which could be overcome should further resources become available.

Discussion

As mentioned earlier slides are collected using consistent lighting conditions at the monastery. Although in most cases these conditions were ideal to capture the range of colours on the manuscripts, occasionally manuscripts have extremely dark covers. Details of such covers are visible on the slide film, but our consistent scanning settings are not suitable for a dark range of colours. Of course, if there were more time available to spend on each slide it would have been possible to improve the colour quality of such images. However, we would still face the problem of recording our adjustments on each image and ensuring that enough technical metadata existed to include in the technical history of the file.

We have been converting our scanned TIFF files to JPEG as well as JPEG2000, since despite the huge benefits of the new file format it is still relatively difficult to find widely used software which supports all the features of JPEG2000. For example, no web browser is natively supporting JPEG2000 and image tiling support is hardly ever implemented by the numerous plug-ins available. Compressed JPEG images are therefore used temporarily for distribution to web browsing software, but we hope that popular web browsing software will soon offer native JPEG2000 support.

Our slide feeder is an efficient piece of equipment considering the bulk of slides to be scanned. Occasionally we find that some of our slides are trapped in the feeding mechanism, delaying the digitization process. Although, it is not clear why certain slides are blocked, there seems to be a connection to the shape of the slide frame and the various grooves and dents present on it. However, this has not affected our progress.

List of tools

Table 1: Software and hardware used in the project, alongside the manufacturer and model

Camera/Lens	Nikon FM2/AF Micro-Nikkor 60mm/f2.8D
Slide film	Kodachrome 64
Colour scale	Kodak colour patch Q-13
Slide scanner	Nikon LS-5000 ED
Slide feeder	Nikon SF-210
Scanning software	Nikon Scan 4.0
Metadata collection utility	In-house
Unique Identifier utility	Microsoft UUID Generator 5.2
JPEG2000 conversion utility	Luratech Lurawave Command Line Tool

References

[1] N. Pickwoad, "The Condition Survey of the Manuscripts in the Monastery of St Catherine on Mount Sinai", The Paper Conservator 28, pg. 33-61. (2004).

[2] PREMIS Working Group, Data Dictionary for Preservation Metadata (OCLC, Dublin, Ohio and RLG, Mountain View, California, 2005).

[3] Technical Standardization Committee on AV & IT Storage Systems and Equipment, Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.2 (JEITA, Tokyo, 2002).

[4] Digital Imaging Group, DIG35 Specification, Metadata for Digital Images, Version 1.1 (I3A, White Plains, NY, 2001).

[5] ISO/IEC, JPEG 2000 Image Coding System: Core Coding System (ISO/IEC, Geneva, 2004) pg. 148.

[6] M. Boliek et al. (ed.), JPEG2000 Part I Final Committee Draft, Version 1.0 (ISO/IEC/JPEG, 2000) pg. 8.

[7] www.hutchcolor.com/PDF/Kodachrome_profiles.pdf (March 2005).

[8] www.marginalsoftware.com/LS2000Notes/casestudyI/scanning_ _kodachrome_on_the_nikon_caseI.htm (March 2005).

[9] IEC, Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space – sRGB, IEC 61966-2-1 (IEC, Geneva, 1999).

Authors' Biographies

Dr Athanasios Velios has studied archaeological conservation at the Technological Educational Institute (TEI) of Athens. He completed his PhD at the Royal College of Arts in London on Computer Application to Conservation. He has been working in digital documentation in conservation for the past 8 years. He has been a lecturer on digital documentation methods in the TEI of Athens. Since 2003 he has been working at the University of the Arts, London/Camberwell College of Arts as a Research Fellow for the Saint Catherine's Library Conservation Project mainly on the digital documentation of Byzantine bookbindings.

Prof. Nicholas Pickwoad trained with Roger Powell and ran his own workshop from 1977 to 1989. He has been Advisor on Book Conservation to the National Trust of Great Britain from 1978, and was editor of volumes 8-13 of the journal "The Paper Conservator". He taught book conservation at Columbia University Library School in New York from 1989 to 1992 and was Chief Conservator in the Harvard University Library from 1992 to 1995. He is now project leader of the Saint Catherine's Library Conservation Project based at the University of the Arts, London/Camberwell College of Arts. He also teaches courses in Europe, America and Australia on the history of European bookbinding.