COLLECTING DIGITAL DATA ON PAPER.

An alternative way for recording conservation survey information.

Dr. Athanasios Velios[1], Prof. Nicholas Pickwoad[2]

SUMMARY
In this paper we briefly describe the methodology followed during the condition survey of the conservation project at the library of St. Catherine's Monastery in Sinai, Egypt. Initially, we focus on the reasons why a condition survey is necessary. Although a computer database is used to store the condition survey data, we explain why the use of computers is not practical for collecting such data on site. We show how computers can be replaced by paper forms and describe the methodology for digitising the forms later in order to transfer the data to the database. We pay particular attention to the way the form should be designed in order for the digitisation to be quick and practical.
KEYWORDS: data, collection, paper, form design, St. Catherine library

1 Introduction

In 2001 the Saint Catherine's Library Project, based at the University of the Arts, London / Camberwell College of Arts, and funded with assistance from the Saint Catherine Foundation was given permission to carry out a condition assessment of the 3307 manuscripts in the library of Saint Catherine's monastery on Mount Sinai, Egypt. The library can reasonably claim to be oldest surviving library in Christendom, and is uniquely rich in undisturbed early bindings. By comparison, no more than 2% of the Byzantine manuscripts in the Vatican Library retain early bindings, whilst in the library of the Athonite monastery of Vatopaidi, no one Greek manuscript escaped rebinding in the nineteenth century. At St. Catherine's, slightly under 50% of the Greek manuscripts (and a higher proportion of those in other languages) retain early bindings. The fact that little was known about the physical make-up of the books within the library, certainly made the idea of a survey more than usually attractive, but it also complicated the development of the survey methodology, as it meant that a definitive process could not be developed in advance of the first visits by the survey teams.

2 Condition survey

The survey is intended to serve two closely linked purposes. Primarily it is to record the condition of the manuscripts so that we can plan appropriate and targeted conservation programmes to deal with the damage that we find, and it will also, of course, serve as a record of the condition of the books at the beginning of the 21st century. In order to do this, however, we have to record the structures and materials of the books so that the risk posed to each book by the damage it has sustained can be properly assessed and the significance of the bindings, in particular, properly understood. The survey therefore combines two elements, a historic record and a condition record, the former attempting, where necessary, to indicate what may now

[1] Research fellow, Camberwell College of Arts, Wilson Annexe, Wilson Rd., SE5 8LU, London, U.K., (a.velios@camberwell.arts.ac.uk)

[2] Project leader, Camberwell College of Arts, Wilson Annexe, Wilson Rd., SE5 8LU, London, U.K., (npickwoad@paston.co.uk)

be missing or obscured by damage in addition to what survives, the latter describing the condition that now presents itself to the surveyors.

However, the remote location of the monastery and the limited time we can have with the books within the monastic day mean that the survey methodology has had to be designed to allow the maximum amount of information to be extracted within the least possible time (Pickwoad, 2004), which we have taken to be an arbitrary average time limit of one hour per book. This means that the survey teams are asked to concentrate on recording raw data, and to leave the analysis of the information for the team members in London, where time, if not unlimited, is at least cheaper and not subject to monastic time restrictions. The analysis of that data is done in London with the help of a specialised computer database, as is normally the case in collections with numerous objects.

2.1 Elements of the paper form

In order to assist in the speed and consistency of data capture, the forms use square check-boxes wherever possible, carefully filled in with pencil to allow for the electronic transfer of the data, as explained later. Text-boxes are also used for recording numerical data or definitions that do not exist in the form already. Finally, drawings are a large part of the forms as described later.

2.2 Layout of the paper form

The pages are divided into numbered *Sections* that lead the surveyor in a predetermined order down each page, each *Section* relating either to different parts of the relevant subject of the page or to the two different types of information (historic or condition) recorded. Within each *Section* there are different *Headings* that relate to the specific pieces of data required, and under each *Heading* there will usually be a column of check boxes with definitions placed next to them. Almost every *Heading* has a check box for *Other*, to be used where the book offers some alternative to the chosen definitions, and a text box next to it in which to enter the new definitions – or a reference to page 10, where space is provided for longer descriptions of what has been found with, perhaps, a drawing. In many cases, more than one check box under each *Heading* may be marked, where more than one of the definitions given might apply.

2.3 Use of drawings

The form also uses outline and diagrammatic drawings to record information, which has the advantage of speed and does not depend on the accurate use of language (English is not the first language of many of the team members). It is particularly suitable for recording areas of damage to parts of the binding such as the covering material, where complex shapes and losses can easily be recorded in outline, but only with great difficulty in words. In addition, the act of drawing by the conservator is a very efficient method of encouraging accurate observation. More specifically drawing on paper (instead of any other medium) allows for better results as explained below.

3 Paper versus computer

Although the conservation survey database is available for the analysis of the data from the condition assessment, all recording is done on paper by the teams of conservators visiting the monastery instead of being recorded directly on the computer. The data is digitized as we will describe later and stored in the database. Here we explain why paper was chosen as a recording medium, instead of using a computer to input the data in the database.

3.1 Familiarity of staff

Conservation training courses to date have not given any emphasis to computer-related applications. As a result most professional conservators are not familiar with modern computer software and its use. On the contrary, emphasis has been given to hand-drawing practice, and many conservators have good drawing skills. Therefore they prefer using pencil and paper than drawing with a mouse (or similar input device) on the computer screen. The advantage of paper is obvious. Forcing the use of computers on site would mean either poor performance by the conservation staff or excluding a great number of otherwise experienced conservators because of their limited familiarity with computer equipment.

3.2 Flexibility of media

The ongoing condition assessment was initially designed based on current (at that time) experience of Byzantine bookbinding. The form allowed for a certain number of features of bookbinding to be recorded based on what had been observed in similar collections in the past. Because of the unique material of the library, many features of bookbinding which had to be recorded, were observed for the first time. The form used for these records could not have been designed to accommodate the newly observed elements. However, because of the use of paper, surveyors were able to modify the form as necessary to accommodate the new features observed.

Similarly, the computer database which could be used instead of paper, could have been modified as well. However, the modification of the database to accommodate the new features would have needed specially-trained staff to work on site as the surveyors would not necessarily have the training to do this work. Because of the remote location of the monastery, employing a member of staff with IT skills to make possible changes in the database, would have been too costly. Remote database development could have been an option, but the reliability and availability of the internet connection at the monastery is rather limited and it would have made the whole process too slow to be practical.

In addition, the flexibility of paper as a medium is not limited to the design of the form. Paper can be used in a variety of ways in which computer input devices cannot be used. A simple example is recording the designs of the impressed tooling on the leather covers of the books. This is traditionally done by placing tracing paper on the embossed pattern and using a soft pencil which is gently rubbed on the paper to reveal the pattern. Such flexibility cannot be offered by any computer input device.

3.3 Practical problems

Other practical problems with the use of computers at the monastery are mentioned here:

1 an unreliable power supply is often a problem at the monastery. Using computers would mean that appropriate UPS was installed, increasing the equipment maintenance cost (there have been cases in the monastery where UPSs have been completely destroyed by unstable power).

2 inappropriate atmosphere. During certain months of the year the wind in the desert can whip up a lot of dust. The safe use of expensive computer equipment would need a controlled environment (this is the case with some expensive photographic equipment currently in use in the monastery), which would again increase the cost and make the condition survey impractical.

3 possible data failure. Using computers at the monastery would mean that all collected data would be stored on site. Because data failure is always likely to

happen, an appropriate backup system would need to be installed. This would increase the practical problems as, once again, it would increase the cost. Moreover, it would have to run without properly trained IT staff and would demand valuable time from the already limited time the teams have in the library.

Having explained the advantages of recording on paper instead of recording directly on a computer, the next section briefly describes the digitization procedure of the paper forms.

4 Digitisation procedure

The digitization procedure consists of two main steps. The first deals with scanning the paper forms. The second deals with translating the scanned pictures to meaningful data which can be stored in the database. We describe each step separately next.

4.1 Scanning

The total number of pages, when the survey is complete, will be about 33000. It is clear that only advanced hardware would be able to cope with such a bulk of material to be digitized. A document scanner has been purchased for that reason. Modern document scanners are equipped with automatic sheet feeders. These allow the scanner to be loaded with a number of pages which can then be scanned as one job, thus minimizing the time needed for placing the originals on the scanner bed. There is a range of available document scanners. Their differences lie mainly in their geometric accuracy.

Here we describe the most common type of geometric distortion observed during the form digitisation with the document scanner:

1 elongation. The document scanner makes use of an automatic sheet feeder. The sheets of paper are fed to the scanner around a rotating cylinder and are scanned in steps relative to the rotation speed of the cylinder. This mechanism introduces some distortion in the resulting image. The images are usually longer than the actual paper pages on the axis along which the page is fed.

2 deformation. The mechanism described in point a) is also responsible for deforming the scanned image. The deformation occurs mainly along the sheet-feeding axis. It is often a combination of elongation in some parts of the page and shrinkage in others.

3 misalignment. The feeding mechanism pulls the paper page by page around the rotating cylinder. However, it is often the case that the page is not pulled completely perpendicular to the cylinder and as a result it is scanned at an angle.

The design guidelines proposed in this document help to correct the problems mentioned above.

Scanning is a rather repetitive process which is almost completely automatic (minor user input is necessary to load the paper and confirm resulting pictures and filenames).

4.2 Inputting

The next step of the digitization process is inputting the data. In the St. Catherine's project this is done in a semi-automatic way. The scanned image of the page is brought onto the screen. It is then read by the software which treats each element of the form individually:

1 check-boxes are recognized automatically as being marked or not.

2 text-boxes are not read automatically due to the limitation of current hand-writing recognition technologies.

3 drawings are transferred "as is".

The user's role is to confirm the readings of the software and type in the hand-written text.

The form can be read automatically by the software if the distortions mentioned in the previous section are corrected. This is because the software holds data for the undistorted page (the ideally scanned image – free from any distortion). In order for the distortion correction to be possible the design of the form must incorporate certain elements which will allow the software to map and reverse the distortion as described next.

5 Design guidelines

In this section we describe some guidelines for designing the survey form in order that it may be corrected (in case any distortions have been introduced during scanning) and successfully read by optical recognition software.

5.1 Registration points

When automatically reading the scanned page, the first step of the process is the registration of the scanned page. The role of the registration process is to match the current coordinates of the elements of the scanned image (distorted page) to the theoretical coordinates of what the image should be (the ideal undistorted page).

The registration is possible by printing small and simple geometric shapes near the edges of the page. The registration process involves the identification of certain points on the page which function as control points. They can be automatically identified by the software based on geometric criteria of these simple printed marks on the page. For example they could be the centre of a circle, or the crossing point of two lines etc. In the St. Catherine's project we propose the use of the innermost corner of a triangle placed near each of the four corners of the page as a registration mark. A simple pixel scanning algorithm identifies the registration point as the last black pixel of the triangle which is closest to the text. This point will always be identified regardless of the distortion or the misalignment of the page as shown in figure 1. Similarly there are triangular marks on each of the other corners of each page.

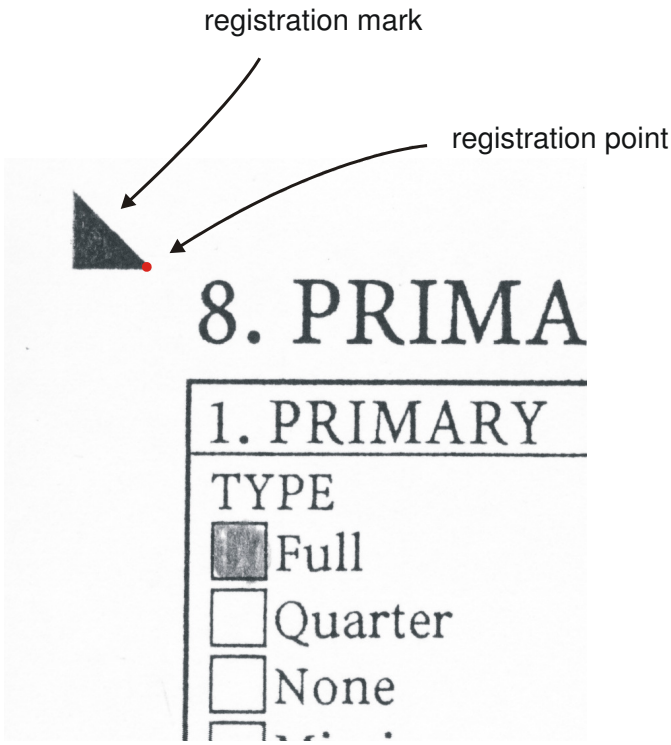

Figure 1: Example of registration point on the printed form

The four points which have been identified on the scanned image correspond to measured coordinates on which the form has been constructed. The distance between the identified pixels on the feeding axis is normally longer than the theoretical distance because of the elongation problem. Such distortion is ignored in the other axis as it is minimal.

By comparing the two distances we can come to a conclusion about how the resulting scanned image can be corrected according to the theoretical coordinates of the perfect page. The way the correction can be done is a rather complex procedure which is not within the scope of this article. However, in order that the correction can be made, registration marks are necessary.

5.2 Check-boxes versus text-boxes

As explained earlier in this article, the form contains a range of elements including check-boxes and text-boxes. It is recommended that check-boxes are used in favour of text-boxes for the following reasons:

1 automatic recognition. Check boxes can be automatically read by the computer software, as it only needs to identify whether the pixels within the boundaries of the check-box are dark or white. On the contrary, automatically reading text-boxes can be a rather difficult process as it depends on the hand-writing of the surveyors, the intensity of the trace left behind (some surveyors use hard pencils) and the currently limited potential of hand-writing recognition software (figure 2).

2 data structuring. Another reason to use check-boxes instead of text-boxes, has to do with the structure of the data recorded. Check-boxes, which correspond to given options on the form, direct the surveyors to record consistently a specific feature using the same term throughout the survey. When text-boxes are used, it is often the case that the same features are given similar but different names by different

surveyors (if a feature of the book needs to be chosen from a list of different options then it is advised that all options are listed next to a separate check-box for each one, rather than including a text-box to write the selected option).

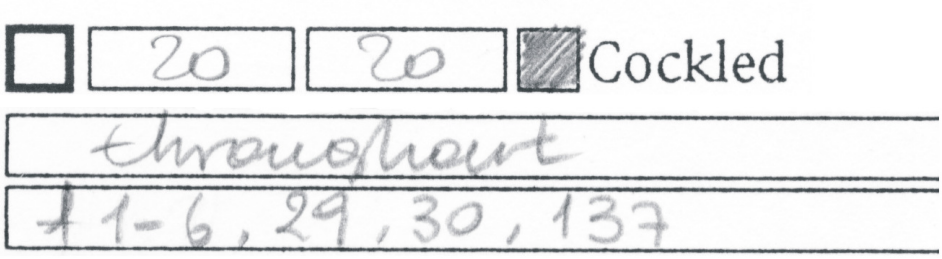

Figure 2: Example illustrating that check-boxes are more easily read than text-boxes.

Check boxes are therefore preferred to text-boxes. However, the arrangement of check-boxes in the page needs to be done carefully as explained next.

5.3 Check-box arrangement

Although check-boxes can be used to maintain the consistency of recorded information, as explained above, there is a risk of data misinterpretation if their arrangement on the page is not planned carefully. A common problem when using check-boxes is that they are placed too near each other in an effort to save space. Two issues need to be emphasised here:

1 unsuccessful registration of the page (which is possible in cases of extreme deformation which cannot be corrected) may lead to check-boxes which are too close together being confused by the software. In such cases, the software would collect the wrong information.

2 careless use of the form may lead to the misinterpretation of the check-boxes. More specifically if the surveyors accidentally mark the area outside the boundaries (figure 3) of a check-box, and this area extends into the neighbouring check-box, then both check-boxes may be misinterpreted as marked although only one of them was intended to be marked.

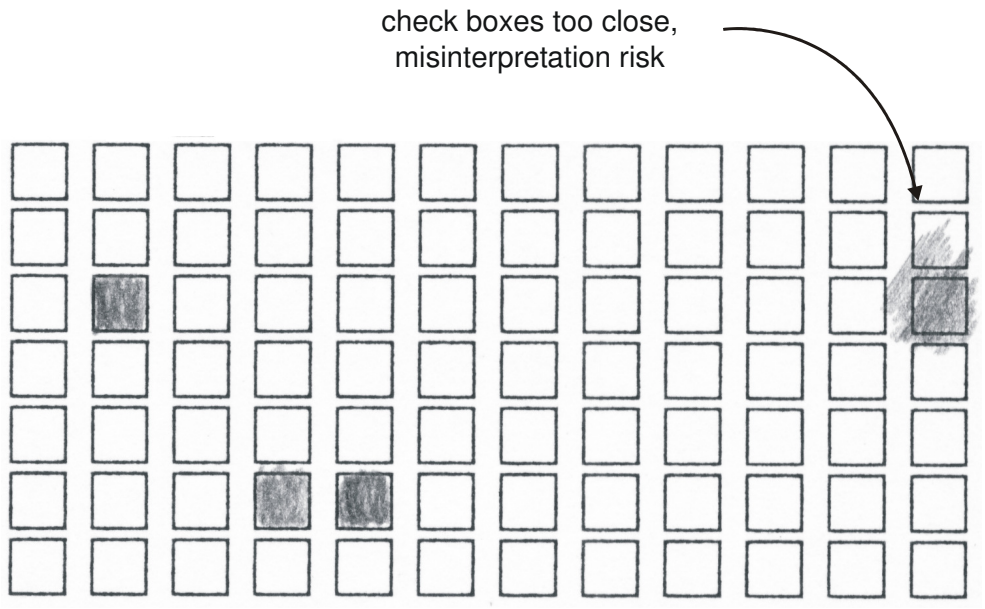

Figure 3: Example of careless use of the form which may lead to misinterpretation

In both cases mentioned above, the problems can be overcome by allowing for adequate space around each check-box. At the St. Catherine's project, the optical recognition software combined with the available document scanner, works adequately with a 1mm space and optimally with a minimum of a 3mm space. The further the check-boxes are away from each other, the less likely it is for the software to confuse them.

Another practice which should be avoided relates to the presence of colour on the forms.

5.4 Colours

When designing a surveying form, complex ideas need to be presented clearly. It is often thought that such ideas can be illustrated by using colour. For example, check-boxes of the same colour are automatically grouped by the human brain and hence can be used to provide an easy way of arranging complex ideas. Although colour is helpful in such cases, its use is not advised here as explained next:

1 reproducing the surveying form in colour (we need one form per item) demands colour printers or printing machines. The cost for such reproductions is significantly larger than black and white reproductions (which can be done on any photocopier).

2 moreover, unless (currently) expensive laser colour printers are used, the time needed to print colour forms with typical ink-jet printers is much greater than printing (and photocopying) black and white forms.

3 scanning in colour results in a much larger file than scanning in greyscale. In the case of St. Catherine's, 33000 colour scans would result is a substantially larger need for storage.

An ideally designed form would, therefore, be in black and white. Any differentiation between elements or grouping of elements should be possible with their correct geometric arrangement. If this is not possible, it is proposed that different line

thickness is used for the boundaries of the check-boxes. Shadowing the background of check-boxes to differentiate them is not advisable as shadowing will be misinterpreted as a marked check-box.

5.5 Modifications

As explained earlier, the survey form for the St. Catherine's project has gone through many revisions to accommodate newly observed features on the books. Although these revisions were initially implemented with pencil on the forms, in more recent trips they have been incorporated into the design of the form. Occasionally these changes demanded a substantial amount of space on the pages of the form and existing elements had to be shifted to make room. As a result, the optical recognition software had to be adjusted to accommodate the new arrangement of the check boxes on the page.

Such problems can be avoided by allowing enough space around elements during the initial design. This way if new elements need to be added, the current ones will not have to be shifted and therefore any modification in the optical recognition software is limited to additions only.

6 Digitisation speed

Having completed half of the condition survey enough data has been collected to give an accurate idea of the time needed for digitizing the paper forms. An indication of these measurements follows (times are per page):

1 scanning takes less than 10 seconds per page (about 2 minutes per form) including feeding the paper to the scanner

2 inputting time depends on the textual information which is present on the form. The more text to be typed in, the more time it takes to input the form. However, for pages with limited textual information inputting takes less than 1 minute, whereas for pages with extensive textual information inputting takes an estimated 2 minutes. The periods mentioned above include the registration procedure which is instantaneous.

When registration marks are not present (i.e. in older versions of the form) automatic registration is unfeasible, and manual registration is necessary. This extends the inputting time by at least 10 seconds (which in total it would mean at least 90 hours of extra work). Also, when complex data is recorded on the form in such a way that it cannot be directly transferred to the database (without following the above guidelines), the time needed for inputting is longer. This is because modifications in the database structure may be needed, or the data may need to be transformed to fit the database structure. Such cases are often observed when new features were identified on the books and it was from the beginning not always clear what kind of data would need to be recorded. This was more often the case on the forms used during the first trips. It is difficult to estimate an average time for these cases but times of up to 30 minutes have been observed.

7 Conclusions

In the previous sections we briefly described the methodology used for the condition survey of the St. Catherine's library. We showed how the use of computer equipment on site is not advisable in our case. The main reasons are the unreliable power supply, the environment and the impracticalities regarding database development and backup on site.

We then explained why paper forms can be used instead of computers for data collection by conservators. The main reasons are the familiarity of conservators with paper and the flexibility of paper as a medium.

Finally, we described how these paper forms can be digitized and transferred to the database, outlining the specific elements of design needed for the digitization procedure to be fast and practical. These are the use of registration points, the preference to check-boxes against text-boxes, the ample space around each element, the lack of colour and the consistency of position throughout the survey for each element. With the times as indicated in the previous section it becomes apparent that the guidelines for designing the form can significantly reduce the time for the digitization process.

8 References

Pickwoad N. (2004), "The condition survey of the manuscripts in the monastery of St. Catherine on Mount Sinai", *The Paper Conservator*, vol.28, p.33.