Collecting and Managing Conservation Survey Data

A methodology for the Saint Catherine's Library Conservation Project at Camberwell College of Arts.

Dr. Athanasios Velios*

Saint Catherine' Library Conservation Project

Camberwell College of Arts

The University of the Arts, London

Wilson Rd., London

SE5 8LU, UK

email: a.velios@gmail.com

Prof. Nicholas Pickwoad

Saint Catherine' Library Conservation Project

Camberwell College of Arts

The University of the Arts, London

Wilson Rd., London

SE5 8LU, UK

email: npickwoad@paston.co.uk

Introduction

With the widespread development of digital technology and digital networks, conservation related data is often "born digital", i.e. it is created in a digital format. A simple example is a conservation database record or a digital photograph. However, a substantial amount of both new and old conservation records are still kept on paper. Paper records offer important benefits to conservators: e.g. they are easily accessible and do not require special viewing equipment. Despite the many benefits of paper, records in this format have a major disadvantage: the reduced capacity for quick searching. This may not appear to be a problem when a record of a single object is examined, but searching becomes difficult when multiple records from a large collection of objects are examined (e.g. to assess the condition of a the whole collection), as the conservator has no convenient way to view the information collectively and is forced to read through multiple pages one by one which is both impractical and time-consuming.

Similar problems were encountered at the Saint Catherine's Project at Camberwell College of Arts. The college undertook a detailed assessment of the condition of the manuscripts from the library of the St Catherine's Monastery in Mount Sinai, during which teams of conservators travelled to Egypt, examined the manuscripts one by one and recorded their observations on paper forms. These forms are extremely useful when studying the binding of a specific volume. Both the binding structure and the condition of a single manuscript are easily retrieved, as each paper form corresponds to a specific book. However, the paper forms are not practical when studying a specific bookbinding characteristic (e.g. bookmarks) collectively. In other words if examination of all different types of bookmarks in the collection is required, paper records are of no practical use as the researcher would need to go through

3,300 pages in different folders to collect the necessary information. In order to allow collective data retrieval the recorded information must be transferred to a computer database system which simplifies searching through large numbers of records.

In this article we describe a novel methodology for transferring paper records to a digital database, so that collective searching is possible. The database was built following the relational model. Its design, implementation and future development using the hierarchical data model have been examined in Velios and Pickwoad (2005 [both]), where the reader is referred to for further information. This article focuses on the paper forms digitisation and data inputting process. We start by investigating a traditional way of computer data inputting based on computer forms and we explain the disadvantages of the traditional model. We continue by proposing a new methodology for data inputting which significantly accelerates the process by partly automating data inputting and partly optimising the computer interface. We conclude with some statistical information about the impact of the new methodology for the data inputting in the Saint Catherine's Project.

The proposed methodology may be of use to other institutions or conservation projects which hold large collections of conservation records on paper and wish to transfer them to a digital database in order to enhance the searching capacity of their resource.

Conservation data recorded on paper

Paper is a convenient format for recording conservation related information as it can be used to keep both hand-written notes (i.e. text and numbers) and drawings. Paper forms are normally used as templates where sets of information are recorded on a per object basis. Such forms are often used in library and archive surveys and as such, it was chosen for the Saint Catherine's condition assessment. The efficacy of template forms is apparent in the assessment of a manuscript. The different binding characteristics are presented in a logical sequence on the form, allowing the conservator to quickly ascertain the condition of the book only by consulting its paper template form and without physically examining the book itself.

Paper form data types

A detailed description of the development of the paper template form used in the St. Catherine's assessment can be found in Pickwoad (2004). The proposed digitisation methodology can be applied to any paper template form, but here we will refer to the Saint Catherine's form to illustrate our points and demonstrate the usefulness of our methodology.

Fig. 1: A sample page from the Saint Catherine's condition assessment form (page 1).

The main characteristic of paper forms is that it combines four types of data (figure 1):

i) hand-written text,

ii) hand-written numbers (and range of numbers "from – to"),

iii) checkboxes and

iv) drawings.

In computing there are different types of data that can be stored in a database. Some of the most frequently used ones are:

i) text (also known as "string")

ii) numbers (there are many types of numerical data – the simplest being "integer")

iii) yes/no (also known as “Boolean”).

Although there appears to be similarities in data types between paper and computer forms, these do not necessarily match each other. In our attempt to transfer data from paper to the digital database, a simplistic approach would be to associate a checkbox on paper (which can be either marked or empty) with a Boolean data type. Similarly hand-written text on paper would be associated with a text field in the digital database (string data type) and a number on paper would be associated with an integer in the digital database. However, such association does not always apply as in many cases a checkbox on paper would act as a shortcut to a textual description and therefore the underlining digital information is textual, despite the fact that a checkbox is used to record it. For example, in figure 2, the marked checkbox refers to the material of a page marker and therefore it answers the question: “What material is the page marker made of?”. It does *not* answer the question: “Is the page marker made of parchment?”. “Parchment” is presented as a checkbox only to speed up the process of filling in the form and it does not mean that the recorded information is Boolean (yes/no). Such design choices for the Saint Catherine’s project paper form were taken into consideration when developing the inputting system as we will describe later in this article.

Fig. 2: Section of the Saint Catherine’s paper form recording the “Page markers material”.

Additional considerations for the paper template form

Before we start discussing the proposed methodology for transferring the paper records to a digital database, it is worth considering some additional aspects of paper template forms with specific reference to the Saint Catherine’s form. These considerations will underline the flexibility of the proposed methodology. The Saint Catherine’s paper form consists of ten A4 pages and six optional A4 pages for features which are not often evident on books – at least in the Saint Catherine’s collection – such as foredge flaps. The form’s 16 pages included more than 1000 fields which need to be checked when data was transferred from paper to digital. Because of the large number of fields our methodology adopted a flexible approach without limitation in the size of the paper form. Although we appreciate that the Saint Catherine’s assessment was completed in a more detailed manner than most condition assessments, our methodology is equally applicable to both extended and compact paper template forms.

In Pickwoad (2004), the authors explain how the paper form evolved during the 5 years of the Saint Catherine’s condition assessment. In the original design it was impossible to predict all types of information that we would encounter during the assessment. The form evolved over the assessment period in order to accommodate new types of information encountered. At the end of the assessment period several versions of each page of the paper form had been used. The various versions of the form, although very similar, were not identical and therefore demanded individual processing. On average among the 16 pages there were about 5 different versions of the form per page. Our methodology was designed to address such inconsistencies.

Finally, on the Saint Catherine’s form, structured text is preferred to free text. This means that whenever text is used, it is devoid of grammatical syntax, rather it is used in the form of keywords to describe a specific feature of or damage to the binding. This is a good approach when it comes to searching, as words are used purely for data and not for grammatical purposes. Therefore, searching is done based on a specific set of terms which are used to describe a feature and it is more efficient than free-text searching which can be arbitrary and return irrelevant results. This is not a limitation of the data transferring methodology but a limitation of the initial method of recording on paper. In general the preference for structured text rather than free text indicates that long blocks of text are rarely evident in the available data.

Traditional inputting on computer forms

Traditionally, data inputting into digital databases is done by skilled typists who read the text on a page and then simultaneously type it onto customised computer inputting forms, similar to the one shown in figure 3. This is a good technique for inputting hand-written free text (paragraphs or notes). Hand-written text is linear and continuous and therefore typing is done in large blocks, which increases the speed of data inputting, especially if the user is a skilled touch typist.

Fig. 3: Example of a standard computer inputting form.

However, in a condition assessment form, data is often stored as structured text, numbers and checkboxes rather than free text. As mentioned above this was the case with the Saint Catherine's paper form. Data in the form is scattered in different areas of a page, thus breaking the linearity of data-inputting. The information is still accessible in a logical sequence on the page, but this sequence does not necessarily follow the arrangement of written text (a line of text under an existing line, etc.). This is illustrated in figure 4 where a few fields of the paper form have been numbered to indicate the sequence with which information should be checked. Therefore, the ability of typing fast is not critical when inputting data from such forms as large blocks of text are rarely encountered. In other condition assessment forms where large blocks of text are used, fast typing would be advantageous. We emphasise that our inputting methodology can be applied to both free-text and structured-text inputting, however, its true potential is evident when a range of structured data is recorded.

Fig. 4: Part of paper form with the sequence of data fields marked on the page.

Checkboxes

A condition assessment form often offers a range of options for a specific characteristic of the object, from which the conservator is asked to choose one. This "one out of many" pattern is usually represented by a sequence of checkboxes on the paper form (figure 5a). This is also the case with computer forms where a control (i.e. a computer interface element) known as "radio button" is used for this purpose (figure 5b). As mentioned earlier, using such controls accelerates the inputting process by eliminating the typing that users have to do and replacing it with a multiple choice. Despite the fact that this process accelerates inputting, the user still has to hit several keystrokes in order to select the appropriate option. This process is also not error free as the keys used for selection (arrow keys) are located next to each other and therefore mistyping is often a problem. The proposed methodology avoids these problems and minimises the margin for mistyping when it comes to data from checkboxes, as we will explain later on, by automatically selecting the correct option.

Fig. 5: "One of many" pattern in paper (a) and digital (b) form.

Drawings

An important characteristic of a condition assessment form, which was also extensively used in the Saint Catherine's form, is the existence of drawings. Drawings are particularly useful when textual descriptions become too long or too complicated. In many cases such textual descriptions can be replaced by a drawing which is completed in seconds and accurately describes the recorded information. Typical examples which illustrate the usefulness of drawings are damage maps, such as the drawing shown in figure 6. Since drawings were deeply integrated in the Saint Catherine's form, it was impossible to exclude them from the data transferring process. Unfortunately, the traditional methodology of manually typing the hand-written information to a text box excluded drawings and therefore a new methodology had to be used for this type of data.

Fig. 6: Example of damage map from the Saint Catherine's condition assessment (page 8).

Having identified the limitations of the traditional approach in data inputting we will now describe our novel methodology for digitising paper conservation records.

Proposed inputting methodology

Scanning

Since drawings are important parts of the recorded data which need to be preserved digitally, capturing the pages of a paper form as images is considered to be a good way of digitising the data. Digital scanners have traditionally been used for digitising sheets of paper. Given that conservation archives can be considerably extensive – the Saint Catherine's condition assessment archive consists of about 33,000 A4 pages and 3,300 A3 pages – methodologies that employ fast scanning equipment are preferred. Therefore, we recommend the use of specialised document scanning equipment which allow groups of complete A4 or A3 pages to be captured in a single scanning job. For the Saint Catherine's project we used an A4 document scanner (ArtixScan 2010) from Microtek and an A3 document scanner (GT 30000) from EPSON. The result of the scanning process is the production of a large number of image files which can be difficult to manage as they all look very similar and they also occupy large amounts of disk space. In the Saint Catherine's project, in order to avoid errors with image management, once the scanning was completed an in-house utility was used to automatically file the resulting scanned images according to the shelfmark of the manuscript that the condition assessment form referred to. Although in many scanning projects the quality of the scanned image is critical, in the Saint Catherine's project, we discovered that standard scanning quality was adequate for our purposes. Therefore, our scanning settings produced black and white images of 300dpi in jpeg format. Although the use of a document scanner was dictated by the existence of drawings, both checkboxes and hand-written text was captured within the scanned page and alongside the drawings. Obviously this data is not pictorial and in the following paragraphs we will explain how the textual, Boolean and numerical data was extracted from the scanned image.

Automatic field identification

Having scanned the paper forms, the next step in our methodology is to read the digital images and extract:

i) textual information,

ii) numerical information and

iii) Boolean information.

As mentioned earlier this information may be located in the form of separate fields (e.g. checkboxes and written notes) in various positions on the page. In order to identify the information stored in these fields, the fields' boundaries must first be identified. This is also true for drawings which rarely occupy a full A4 page: when examining a drawing, a user will only want to see the specific part of the page rather than the whole sheet. Therefore the boundaries of a drawing within a page also need to be identified.

Fig. 7: Logical diagram of the utility for automatically identifying information on the paper form.

We have developed a software utility which is capable of automatically identifying the boundaries of the fields of a page where each piece of data is located. The logical path that the utility follows in order to succeed can be seen in figure 7. The utility loads the image of a specific page and registers the location of each field on the form using the coordinates of the field on the page (x: distance from the left edge of the paper and y: distance from the top edge of the paper, figure 8). Because there may be hundreds of fields in a single page, the coordinates are kept in a spreadsheet where they can be managed using specialised spreadsheet software. Having registered the locations of all fields of a page in the software, a rectangular box is overlaid on top of the scanned image of the page to guide the user to a specific field (figure 8). The user is then able to navigate through the different fields using the keyboard.

Fig. 8: Example of identified checkbox on the scanned page.

Although this methodology generally works well, we need to emphasise that the registration of the fields on the scanned image can only be successful if the page has been scanned correctly, i.e. it has not been accidentally cropped or resized during scanning. We underline this point, as low-end scanning equipment can introduce a degree of geometric deformation in the long dimension of the page which is due to the paper feeding mechanism of the scanner. Regardless of the experience of a user, when multiple sheets of paper are fed into the scanner, accidental cropping and misalignment of the page can occur. Such limitations of the hardware cannot be entirely prevented but they can be corrected using manual registration of the fields on the scanned image. Therefore, the coordinates recorded for every field can be measured from points on each page which should be away from the page edges and thus guaranteed not to be accidentally cropped during scanning. Moreover, in order to take into account any geometric deformation of the scanned image, two registration points need to be manually identified on each page. The in-between distance of the points on the paper page can be measured and compared to the distance of the points in the deformed digital image. This comparison can then be used to correct the deformation of the scanned image and successfully register the fields.

Fig. 9: Example of a registration mark on the Saint Catherine’s paper forms.

In the Saint Catherine’s project, the forms have been equipped with triangular registration marks (figure 9) at the corners of each page (figure 1). These allow the registration to be performed automatically, as the software can identify the distinct shape of the marks, automatically measure the distance between them and compare that distance with the real distance measured on paper.

Having explained how the location of each field is identified on the page, it is now appropriate to describe how the data inside each field is treated by our software.

Minimal interface

An important characteristic of our software is the fact that it lacks any ordinary computer interface controls (such as radio buttons, text boxes, choice menus etc.). We avoided using such controls as they can slow down the inputting process since users prefer using the mouse when working with such controls, which is significantly slower than working with the keyboard. Therefore, we minimised the interface to only the scanned image and the overlaid rectangular box and we removed any computer interface controls which would otherwise slow down the process. Instead of requesting the user to actively input data, the software initiates the inputting process and requests the user’s confirmation before saving the data. This initiative is different according to the type of data examined as we describe next.

If the overlaid box marks a *checkbox* then the software will automatically examine whether the checkbox is filled in or blank. If the checkbox is blank the software assumes that no information is stored in that field and automatically moves to the next one. If, however, the checkbox is filled in, the software will recommend the corresponding (often textual) value to the user. At this point the user can either agree to the software recommendation or skip the field, if it has been wrongly identified as “filled in” by the software.

A similar approach is followed when fields for *hand-written text* exist on the form. If the field has been left empty then the software ignores it and moves to the next one. If however, the field is filled in, then the software requests the user to manually type in the hand-written text. We do not use hand-writing recognition software, although it has been improving over the recent years, as we felt that it is still too difficult to incorporate into our utility. In the Saint Catherine's project, more than 30 conservators took part in the assessment and training the hand-writing recognition software to understand all the different hand-writing styles would have been a time-consuming task which we chose to avoid.

Similarly for *drawings* the software stops and awaits confirmation by the user on whether the specific drawing needs to be stored as a separate piece of data or not. It is worth mentioning here that the boundaries of small drawings in a page are given by:

i) the coordinates (x, y) of the drawing field as well as

ii) the width and height of the drawing (w, h).

All boundary values are stored in the spreadsheet. When a user confirms that a drawing needs to be saved, the software saves the boundaries of the drawing without creating a new image. Therefore, in order to recall the drawing at a later stage, it is only necessary to crop the original scanned page using the boundary information of the drawing (this information may vary from page to page because of the geometric variations introduced during scanning). Saving drawings using this technique is economical in terms of disk space, as only one file is used for all drawings on a page and therefore information is not duplicated in multiple cropped images.

The above procedures indicate the simplicity of the software, whereby the user has two choices at any given time: either to agree with the software's recommendation and/or type the necessary data or to disagree and skip the field to go to the next if this has been accidentally identified. The simplicity of the software means that the training period for a potential user is minimal and the data inputting is done very quickly. Some statistics about the speed with which data can be transferred is included at the end of this article.

Directed user focus

In trying to further reduce the inputting time, we have introduced another feature in our software which augments the presentation of the fields and the scanned page on screen. Traditional interfaces heavily rely on scrolling a page or a form to reveal the controls hidden below the bottom of the screen, as often screens are too small. This has two disadvantages:

i) scrolling slows down the process while distracting the user and

ii) usually the point of focus for the user is at the bottom of the screen where the hidden controls have just been revealed; we explain next why this is a disadvantage.

The fact that the focus is kept low on the screen means that a large part of the form is always hidden and although viewing the whole form is not critical for a user, knowing where the next field will appear is certainly helpful to the workflow. Moreover, low focus screen point is not ergonomically advisable for long period computer users. Therefore, we decided not to introduce any scrolling in our interface but instead kept the focus of the user to a fixed point on the screen and moved the digitised image of the form instead (figure 10). This meant that the user did not have to constantly focus on different parts of the form which further increased productivity.

Fig. 10: The point of focus remains the same while the fields change one after the other.

Keyboard use

Having optimised the workflow for speed of data inputting, we tested the use of the software and we discovered that our approach significantly reduces the data inputting time, as we will explain in the following section. Because of this optimisation we observed that during testing there was an increased risk for errors due to the proximity of the designated keys on the keyboard for accepting or rejecting values, which inevitably led to the wrong keys being pressed. For this reason we reassigned the functions on other keys which were located on different sides of the keyboard. Therefore, the “Tab” key on the left-hand side of the keyboard is used for skipping fields (rejecting recommendations) combined with the left “Shift” key for backward navigation. The “Enter” key on the right-hand side of the keyboard is used to accept a value. The numerical keypad on the right-hand side is used for number inputting. Finally, text is typed using both hands. This function/key rearrangement clearly separates the roles of the two hands, with the left being responsible for navigation and the right for confirmation and numeral typing, thus reducing the risk of errors due to mistyping.

Results and short conclusions

The proposed methodology for transferring data from a paper form to a digital database focuses on minimising the risk for mistakes and reducing the transferring time. This methodology was developed for the digitisation of the condition assessment data collected during the Saint Catherine’s project. A computer form based on traditional computer controls (e.g. radio buttons or text boxes) was tested. The time needed for inputting a single page of the paper form was between 60 to 90 seconds and that methodology excluded drawings. Our methodology reduced that time to between 20 to 30 seconds including drawings. This reduction in inputting time is particularly important when many thousands of pages need to be digitised. In our case one person managed to complete the transferring of data from about 33,000 pages within a year, which would have been impossible using a traditional computer inputting form.

Therefore, our methodology has been applied to the Saint Catherine’s project with satisfying results. The same methodology can be applied to any digitisation project for conservation archives as the software is fully customisable to individual needs. We are keen to apply our methodology to other case studies and if necessary to further optimise our software.

References

Pickwoad, N., 2004, *The condition survey of the manuscripts in the monastery of St Catherine on Mount Sinai*, The Paper Conservator, Volume 28.

Velios A., Pickwoad N., 2005, *The Database of the St. Catherine's Library Conservation Project in Sinai, Egypt*, IS&T Archiving 2005 Conference, April 26-29, 26/04/2005, Washington, DC, USA.

Velios, A.; Pickwoad, N., 2005, *Current use and future development of the database of the St. Catherine's Library Conservation Project*, The Paper Conservator, Volume 29, p.39-53.