Navigating Through Uncharted Territory:IDP, An International Internet Digitisation Project

[Paper presented to the London Advisory Committee of the British Library in 2001, previously unpublished]

Dr. Susan Whitfield, The International Dunhuang Project

Summary

Digitisation is:

- Labour intensive
- Expensive
- Requires considerable space and resources
- Requires staff with expertise in areas not traditionally found in libraries/museums etc.
- Requires long-term maintenance to avoid obsolesence

It should therefore not be considered unless:

- It adds considerable and relevant value to the materials under consideration
- Staff, space and resources are available
- Provision for long-term maintenance has been considered and can be provided for

Contents

The discovery at the turn of the last century of a hoard of Chinese documents in a cave in Chinese Central Asia was unprecedented, certainly in Asian scholarship. Not only were there a vast number of documents in the find - 40,000 as a round figure - but many are primary historical documents not found anywhere else and, moreover, totally different in kind from the usual Chinese historical sources with their concentration on Buddhism and glimpses into ordinary - rather than court - life. Within 15 years of their discovery the documents were dispersed to institutions in London, St. Peterburg, Paris and Beijing and have never been completely accessible to scholars since. Yet their study is not only of interest to Chinese historians: they also shed light on one of the most important regions for world history - Central Asia - containing unique information about its many civilisations and cultures during the first millennium AD; as well as being important for the history of religion; the development of paper making and printing (the world's earliest dated printed book (AD 868) was among the documents); and the evolution of the book. The International Dunhuang Project (IDP), founded in late 1993, is a collaborative endeavour to ensure the best preservation of, and increased access to, these documents. This paper, after introducing the history of the documents, will discuss some of the lessons learned by IDP in the course of development of its Interactive Web Database.

RETURN TO TOP

The History of the Documents

Dunhuang, a town now in the Chinese province of Gansu, is on the eastern part of the Silk Road. Throughout the first millennium AD, monks, merchants, soldiers and diplomats travelled this route from one thriving oasis to another finding them to be sizeable towns with inns, markets and irrigation systems supporting considerable agriculture. There were also Buddhist monasteries and large garrisons. But climate changes and invasions forced the people of these communities to abandon their once fertile lands in the latter half of the first millennium of the common era and in time the sands of the deserts covered their homes, the Buddhist shrines and the garrisons.

The first decades of this century saw great Western archaeological interest in this area founded, it must be said, less on co-operation than competition. Among these archaeologists were Aurel Stein, a British national of Hungarian descent; Paul Pelliot, a Frenchman; and Sergei Oldenburg, a Russian. It was these three who made their way east, having heard of a remarkable cache of documents. In 1900 a sealed-up and hidden cave was discovered near Dunhuang stuffed floor to ceiling with manuscripts, printed documents and paintings. The three carted away roughly a quarter of the cache each to London, Paris and St. Petersburg respectively where they remain today. Most of the remainder were taken to Beijing. There are also a few hundred in Japan from the expeditions sponsored by Count Otani and smaller numbers in institutes and private collections throughout the world.

RETURN TO TOP

Twentieth Century Conservation, Cataloguing and Collaboration

Although many of the documents were in excellent condition when they arrived at the various institutions, the large quantities coupled with the considerable disruption of two world wars, has meant that none of the institutions has completed conservation. Moreover, the uniqueness of the find and the lack of experience in Chinese paper and book forms in most of the institutions resulted much inappropriate early conservation work which must now be reversed.

Cataloguing has also been tardy. None of the collections are completely catalogued. A catalogue of 7,000 items in the British Library Stein collection was published in 1957 but this omitted much fragmentary material. In the 1960s and 1970s the British Library produced microfilms of large parts of their collection and other institutions followed. Pirate facsimile publications, taken from these microfilms, were produced in Taiwan in the 1970s, giving scholars much greater access although the quality was very poor.

The late 1970s and 1980s saw the start of collaboration between scholars and conservators worldwide. There were several exchanges between the National Library of China and the British Library during this period and over 5,000 fragments were duly conserved. The non-Buddhist material was published in China in 14 volumes in a Sino-British project, and two Chinese scholars are currently working on cataloguing the fragmentary material.

In 1993 Peter Lawson, then Head of the Oriental Conservation Studio at the British Library, organised a conference in the UK. For the first time, conservators and curators from all the major holding institutions met to discuss this material. The conference delegates were keen to start working together to learn from each institution's mistakes and successes. Since then similar conferences have been held at different venues every two years (Paris, Berlin, St. Petersburg and, in 2001 and 2003, Stockholm and Beijing).

RETURN TO TOP

The International Dunhuang Project

At the close of the conference it was decided to form a collaborative project, later named 'The International Dunhuang Project', 'to promote the study and preservation of the Dunhuang legacy through international collaboration.' As the result of a successful application to the Chiang Ching-Kuo Foundation for International Scholarly Exchange in Taiwan, Dr Susan Whitfield was employed to run the Project from the British Library: to publish a newsletter (thrice-yearly), to organise the conferences; publicise the Project, and to fulfill the Project's aims by developing a database containing full information about, and images of, the documents, which database would eventually be made available on the internet.

The database was developed by Dr Whitfield with two year's free consultancy from the software company, and data was inputted over the following three years. In October 1998 the database went live on the Internet with entries for over 20,000 items in the British Library collection and over 2,000 images. As of October 2000 it contains details of over 28,000 items and over 15,000 images. A new map interface was launched in June 2000, taking advantage of the wealth of Stein maps, photographs and site plans held at the British Library.

IDP is recognized as among the leading digitisation projects in the art and humanities field (dossier available) and consistently places top among search engines for queries asking for Dunhuang. A five year joint project with the National Library of China, Mellon Foundation funding for a four year project, and other funding and collaborative agreements currently under discussion will expand and accelerate IDPs work over the coming years, ensuring its place in the very top rank of digital projects worldwide.

This has been achieved on a budget of between 20-60K per annum. From 1993 to 1998 the Project had only one-full time staff member (plus one part-time data inputter and a database consultant). It now has three full-time staff. It has been consistently successful in raising external funds and sponsorship and, up to April 2000, all staff members were funded externally.

IDP staff are regularly asked to demonstrate the Project to BL visitors, from the Treasury to the Chinese Minister of Culture, and are also regularly asked on an informal basis by BL staff for advice on digitisation, equipment, standards and contacts in the field. The BL has not given IDP any institutional guidelines or standards to follow and has offered no advice itself on equipment or methodology. These have been gained from top projects around the world which have been generous with their time and expertise. However, IDP is an innovative project and is therefore leading the way rather than following: it has largely had to make its own charts for this previously unexplored territory.

IDP staff have not been invited to contribute their expertise and experience to the BL Digital Library Project, although they are members of technical groups and committees for several international projects and are approached with regularity by scholars from institutions worldwide to offer advice, give demonstrations and lectures.

RETURN TO TOP

The Case for Digitisation

The International Dunhuang Project is a compelling example of how computer technology can be applied to make resources available in a manner previously unimaginable. Even if scholars travel to all the places where the manuscripts are held and buy all the catalogues, microfilms and other facsimile forms now available, they will not be able to see the greater part of the collection.

The main reason for the limited conservation and cataloguing has been - and continues to be - a combination of lack of resources coupled with lack of sufficient numbers of scholars and conservators with the necessary expertise. This situation will continue to apply nationally in the foreseeable future. In other words, working in isolation it is unlikely that any of the major collections will complete conservation and cataloguing and open their entire collection to scholars. Even if they were to do so, the preservation of the manuscripts would continue to be a limiting factor to their accessibility. Some are so frail as to make any handling inadvisable; and even though the majority are in good condition, handling should still be kept to a minimum to ensure their longevity. Moreover, each collection is still only part of a whole which needs to be available in its entirety to be understood fully.

Therefore there is a compelling argument for international co-operation to create a widely available computer catalogue of all the manuscripts with images. It is the only means to provide scholars with access to the entire collection. Using images means that the manuscripts can be studied both by scholars interested in the text and those interested in the object. Moreover, despite the massive input of resources needed to achieve this end, it will save each individual institution considerable work because of the exchange of ideas and the possibility of sharing techniques, especially in conservation and catalogue design. They can also combine their energies for fund-raising and the collaborative nature of the Project should increase the chances of success in this vital area. The Project will also make the manuscripts available for research to many more scholars whose work will contribute to the maintenance and updating of the database.

Because of the differing physical locations and the age and condition of the manuscripts, it is vital that, if they are to be made accessible to more scholars, it is achieved mainly by using surrogates. Providing transcriptions of the texts would not be satisfactory for several reasons. Firstly, the texts are often illegible or at least open to different interpretations. Second, they contain many characters and glyphs that are not found in existing Chinese character coding systems. Third, a full text transcription does not show the context, handwriting, type of manuscript, colour of the paper, and numerous other details that could be vital for understanding and dating the text. This being said, it is planned that transcriptions will be included in the database or, in most cases, links made to existing electronic resources containing transcriptions such as, for example, the Buddhist Canon. IDP is about to start a collaboration with a Buddhist institute in Taiwan to achieve this.

Other forms of providing a complete catalogue with images would be so unwieldy and expensive as to make them effectively inaccessible to many scholars. The Internet is now accessible to most scholars worldwide and image-download time is improving year by year.

Egmond's argument that a computer catalogue must fulfil four basic characteristics to rival its printed form, namely completeness, flexibility, usability and portability, is accepted (with the caveat that any printed catalogue of the Dunhuang collection would be far from portable (Egmond 1990)).

RETURN TO TOP

The Perils of Digitisation

Too often libraries, museums and other institutions do not consider whether to computerise but only when. The costs of computerisation are high, both in setting up (for equipment, technical support and man-hours) and in maintenance, and these need to be weighed against the expected benefits. For an example Hahn (1990) has estimated that an average of 1.1 hours is required per manuscript for encoding, proof-reading and making corrections, and this does not include time for producing digital images.

Despite the fact that computer technology is not new, the skills and knowledge required to exploit its potential and the infrastructure of backup and services are still patchy. In 1991 W.V. Egmond observed that computers were being used much as printing technology was in fifteenth century Europe: to provide an alternative means of doing what was already being done. Nine years later, technology is still often used unnecessarily, rarely fully exploited, and has brought new problems which are sometimes not acknowledged: it is time-consuming, expensive and, because of incompatibility, in some cases may prove to be of little long-term use. Printed catalogues have served very well and extra time may be best spent, for example, in providing a concordance and improving the index rather than on computerisation. There should be a strong argument for the additional benefits of computerisation which will counter the additional costs.

RETURN TO TOP

Avoiding the Perils

Over the past eight years I have encountered a thousand or more digitisation projects in the arts, humanities and social science fields. Of these, I estimate that fewer than 50 have been well-conceived and fewer than 20 well implemented. Most of these thousand projects are already obselete or will become so within ten years. Yet these thousand projects have commanded vast resources, most of which could have been better deployed in more traditional scholarship, preservation or publication. Many of them have been carried out by internationally recognized institutions or by major scholars in their fields. It is now time that key institutions, such as The British Library, lead the way by preparing strict guidelines for choosing, designing, implementing and maintaining such projects so that further precious resources are not wasted. There is usually no need to commission reports from outside experts to draw up these guidelines: the expertise and knowledge is available freely, it simply needs to be pulled together and implemented consistently across the library.
RETURN TO TOP

To summarize:

- *1. Choose Projects Carefully*
  Projects need to be well selected and open to peer review at the planning stage so that problems are identified before resources are spent. Most projects I see in the course of my work are ill-conceived and should not have been started. The main shortcomings are:

  - **Digitisation adds no or little value to the collection**
    e.g. it is a small, already accessible collection of interest to a limited audience and with illustrated, published catalogues and the scope of digitisation is the same as that of a published catalogue.

  - **The project is concerned mainly with producing pretty pictures without content or context.**
    Anybody can mount a few images on the web. The advantage of the BL is that, first,it has unique resources of international significance and, secondly, it has specialists who know the material.

  - **The project replicates existing work.**
    e.g. there are at least three projects to digitise the Pali Buddhist canon, two for the Korean canon etc.

  - **The project allows for no possibility of expansion.**
    e.g., to make it compatable, for example, with common interfaces (ECAI, for example), or with other, similar databases.

  - **The project does not aspire to minimum standards.**
    e.g. digital images are produced without colour bars, measures etc. and at too low quality.

  - **The project does not exploit digital technology to the full.**
    This is a shortcoming almost universal among digital projects. While it will probably take a new generation to realise the potential of computerisation and the Internet, we should still try to avoid simply replicating the old technologies.

- *2. Select the projects before applying for funds*

- *3. Ensure that sufficient and continued resources are available and that they are best spent on this work*

- *4. Plan for space and hidden costs*
  Photographer needs dark room space. Digitisers need natural light and a place where they can interact. Server computers are usually required. There may be a storage cost to build in. Conservation is likely to be a major cost.

- *5. Plan for long-term maintenance of the digitial data*

- *6. Learn from others*
  There is a tremendous amount of expertise in the community. From the beginning IDP has made contact with people working on similar projects, surveyed the literature, evaluated existing technologies, methodologies and standards and, on the basis of this, carried out its own pilot studies to take account of the special nature of the material to be digitised. We now share this expertise - including the database structure, the pilot studies and other written reports - with others worldwide. It would be helpful if there were a central BL body to keep up with changes in the field, collate the information and make it available to BL project staff.

- *7. Use existing and common standards*
  Eg. TIFF, Dublin Core, Unicode etc.

- *8. Use well-established hardware and software where possible*

- *9. Use curators' expertise but do not expect them to become digitisation experts*
  i> Their expertise is what makes the British Library and other such institutions unique, but, as scholars, they do not necessarily have the skills for photography, image manipulation, database and web page design and the numerous other skills required for digital projects. Nor can it be assumed that they can be trained in these skills.

- *10. Maintain standards and avoid gimmicks*
  The BL does not want to be seen to be producing sub-standard work: using cheap printing techniques or semi-literate editors would not be considered in BL book publication: no different for digital publications.

- *11. Think Ahead & Build in Flexibility*
  Image downloads/GIS/Unicode etc.

RETURN TO TOP

Bibliography

Egmond, W. V., 'Principles for Computer Catalog Descriptions of Medieval Scientific Manuscripts', in Folkerts and KŸhne 1990, pp. 109-122.
Egmond, W. V., 'The Future of Manuscript Cataloguing', in Stevens 1991, pp. 153-158.
Faulhaber, Charles B., 'Philobiblion: Problems and Solutions in a Relational Database of Medieval Texts', Literary and Linguistic Computing 6.2 (1991), pp. 89-96.
Folkerts, M. and KŸhne, A. (eds.), The Use of Computers in Cataloguing Medieval and Renaissance Manuscripts (Algroismus: Studien zurr Geschichte der Mathematik und der Naturwissenschaften: 4), Institut fŸr Geschichte der Naturwissenschaften, Munich 1990.
Giles, Lionel, Descriptive Catalogue of the Chinese Manuscripts from Tunhuang in the British Museum,

The Trustees of the British Museum, London, 1957.

Hahn, N. L., 'The Future of Computerised Manuscript Catalogues. A Proposal', in Folkerts and KŸhne 1990, pp. 41-56.

Jefcoate, Graham, 'Getting the Message from Gabriel', Library Technology 1.2 (April 1996), pp.33-34.

Mayo, Hope, 'Standards for Description, Indexing, and Retrieval in Computerized Catalogs of Medieval Manuscripts', in Folkerts and KŸhne 1990, pp. 19-40.

Mayo, Hope, 'MARC Cataloguing for Medieval Manuscripts: An Evaluation', in Stevens 1991, pp. 93-152.

Muller, Charles,

Stevens, W. M., Bibliographic Access to Medieval and Renaissance Manuscripts: A Survey of Computerized Data Bases and Information Services (Primary Sources and Original Works: 1 (3/4)), Haworth Press, New York 1991.

Stevens, W. M., 'Sources and Resources for History of Science to 1600: A Survey of Computer-assisted Catalogues for Original Sources in Manuscripts', Nuncius: Annali Di Storia Della Scienza 9.1, pp. 239-264.

Whitfield, Susan and Wood, Frances (eds.), Dunhuang and Turfan (British Library Studies in Conservation and Science: 1), The British Library, London 1996.

RETURN TO TOP