THE HIGHER EDUCATION DIGITISATION SERVICE: MANAGING THE CONVERSION TO ELECTRONIC FORMATS

Simon Tanner

This article will introduce HEDS and its services, with the focus on its serials related projects and the issues that are raised by digitising serials.

Simon Tanner, Digitisation Consultant, Higher Education Digitisation Service, University of Hertfordshire. S.G.Tanner@herts.ac.uk

Introduction

The aim of the Higher Education Digitisation Service (HEDS) is to establish a range of core and value-added services available through a single point of contact to support the conversion of high volumes of learning, teaching and scholarly materials into electronic forms for increased availability in the higher education community.

HEDS has been established by the University of Hertfordshire by bringing together a wide range of expertise and specialisms with project funding from the eLib programme. This project funding was initially granted for 2 years from September 1996. There are 3 key participants.. The University of Hertfordshire provides the management of the project and the staff who offer all the advice, consultancy and copyright research services to our clients. Cimtech Limited, a wholly owned subsidiary of the University, has over 30 years experience in document management and is contributing consultancy and technical support to HEDS. International Imaging Limited, an independent commercial operation associated with the Chadwyck Healey Group is providing the 'engine room' for the service, using HEDS equipment to carry out timely, quality conversion work to HEDS technical specifications.

The HEDS project plan has three phases: to establish and then offer digitisation services, followed by phased transition to a self supporting service. In Year 1, from September 1996, HEDS was established, equipped and staffed. A number of digitisation jobs, also funded by eLib, are progressing well to 'pipeclean' the HEDS processes and services. HEDS has been able to establish effective business processes and test our technical capabilities through these projects. In Year 2, HEDS is now open for business and ready to accept digitisation work by request from a range of clients, both HEIs and non-HEIs. In Year 3, HEDS will begin the transition to a self supporting service through the implementation of a business plan agreed with the JISC.

Services

The main objective in the HEDS service plan is to provide a total management package so that our clients have a single point of access to a range of co-ordinated and interlocking services. HEDS will work with clients to agree and deliver a complete package, including all the issues identified in Figure 1 below.

At this stage in the service development, HEDS is focusing on the following core services.

- advice and consultancy to clients on the feasibility of digitising defined collections of materials;
- guidance on selecting the most cost-effective methods for realising your digitisation aims;
- to manage the complete job, from problem definition to final product delivery and acceptance;
- to prepare functional and technical conversion specification for the digitisation work;
- to provide quality assurance procedures to validate the end product;
- to deliver digitised materials on time and to agreed standards;
- HEDS will also provide advice and assistance with copyright clearance;
- HEDS will take a central role in raising awareness of digitisation for the higher education sector.

HEDS sees its role as being able to advise and carry out production to support all of the elements identified in Figure 1 and especially to extend the specialist services. HEDS has drawn together extensive experience in handling valuable, unique and fragile materials with appropriate care and security. One of the obvious benefits of creating electronic versions of such materials is the preservation through alternative access reducing wear and tear on the original. HEDS is also already providing copyright research in connection with some projects. Further development will support the provision of copyright clearance services and referral to expert legal advice, where appropriate, as value-added services.

Projects

HEDS has a number of eLib funded projects in progress at the moment which will result in much wider access to the materials with more than 105,000 pages converted. These projects include the following materials:

- Archaeology research reports
- *Transactions of the Institute of British Geographers*
- Meteorological observatory data (1881-1975)
- The Statistical Accounts for Scotland (1799 & 1845)
- Social policy and transport pamphlet collection of the British Library of Political and Economic Science

These will present a valuable resource to the HE community and a number of technical challenges for HEDS in converting a range of materials into various electronic formats. The original materials include 35mm microfilm, paper sizes from A5 to greater than A3 with handwriting, printed text and graphics including photographs. The electronic output has required techniques including image scanning, optical character recognition, conversion to Adobe Acrobat PDF and rekeying of some indexes. In the following section I will describe in more detail some of our more serials related projects that will

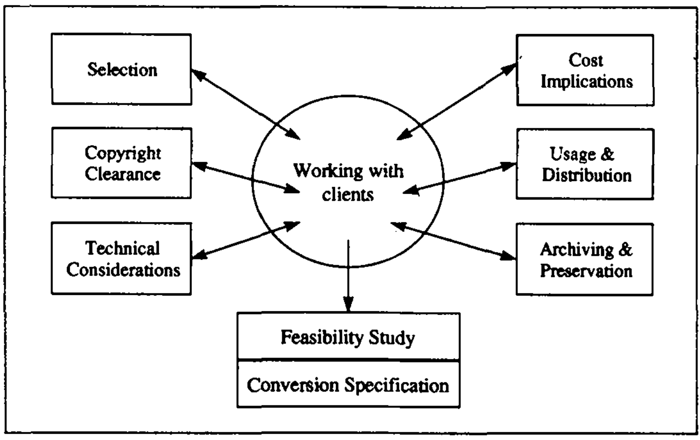

Figure 1: HEDS Service Development

develop further the issues of digitising journals for use in HE.

Transactions of the Institute of British Geographers

The *Transactions of the Institute of British Geographers* is a journal spanning the humanities, social sciences and natural sciences, produced by a scholarly society (Royal Geographical Society with the Institute of British Geographers). This journal runs from 1935 and HEDS is digitising the whole backrun for free access by UK HE. Early samples of material converted by HEDS may be found at the client web site[1].

HEDS worked with the Parallel Publishing for Transactions (PPT) eLib project[2] to make an initial assessment of the materials, the digitisation issues and the preferred way to present the information. As the PPT had already served a recent issue of the journal in PDF format, there was plenty of evidence to demonstrate that converting the backrun into PDF would be the most effective technology for its needs. However, because of the cost implications for conversion of the early editions, which have poorer quality originals, these will be converted to 'image and text' format PDF whereas 1970 onwards will be in fully converted PDF.

Nature

HEDS is about to take part in the pilot study for the complete digitisation of *Nature* (1869 to the present day). This pilot is funded by JISC's Committee on Electronic Information. Macmillan Publishers Ltd will make the source material available. Initially, there will be 30,000 pages converted and made available to HE as part of an evaluation. HEDS will determine the costs of digitisation to an acceptable quality and meet the challenges that are likely to arise when digitising the different fonts, styles and paper that Nature has utilised over the past 130 years. This pilot is in the very early stages, but is expected to follow the general model established by JSTOR[3] of images for each page viewed on screen with hidden searchable text. The initial 30,000 pages will follow topics chosen by the publishers, Macmillan.

World's Fair Newspaper

The University of Sheffield holds the National Fairground Archive and as one aspect of a grant from the Heritage Lottery Fund they are submitting a plan for the digitisation of the *World's Fair* newspaper[4]. HEDS is providing consultancy on the options available for digitisation. The *World's Fair* newspaper is the main trade organ of the fairground community. Since its foundation in 1904 it has been published weekly and the total volume would approximate to 350,000 broadsheet sized pages. There are paper originals available and also microfilm copies. This project will require balancing technical issues of matching image clarity and retrieval needs against cost benefits and end user requirements.

Issues raised by digitising journals

The format and layout of serials have many ramifications for their digitisation and the electronic outputs that can be achieved. Most of these are pragmatic considerations based on the cost benefit ratios of differing approaches to the material and the target application for serving the end product.

There is a wide range of page formats to be dealt with from A4 size to full broadsheet newspaper sizes. In some cases the serial publication may only be available on microfilm for digitisation. The paper and print quality can both have a distinct effect on the techniques used for digitisation to present the best end product.

These factors may vary considerably over the lifetime of a single title and therefore no single approach should be considered definitive for the whole collection, although there is a range of best practice that can identified. Newspaper formats have been approached in a number of ways from first microfilming the newspaper and then digitising from the microfilm or coping with the large format originals by using a bookscanner. HEDS prefers the latter approach as it gives greater control over the image quality and text reproduction. For more standard sizes, bookscanners speed up the scanning process when dealing with journals in large heavy bindings or where stripping or disbinding is not an option.

The layout of journal pages is generally required to be reproduced as closely as possible to the original because of the various rights and agreements required to digitise. This means that only a few electronic formats will be acceptable as the means of fully representing the layout. Primary formats are a straight image file such as TIFF or the Adobe PDF format that allows for character recognition while retaining the look and feel of the original. Whilst HTML or SGML marked up data might be more effective as a means of transferring the information to the user in smaller file sizes, the problem of representing the originals look and feel often hinders their use and also adds significantly to the cost of conversion.

In addition to layout, content accuracy is also an issue. Optical character recognition (OCR) is not perfect and errors in the text can occur. This is particularly likely where there are large amounts of tabular information, diacritics or scientific notation. Where OCR is carried out then the error levels should be less than 1 in 10,000 characters. Anything below that level falls below most original publishers' printing specifications. In certain cases, for example Nature, accuracy is an issue of utmost importance and no error would be acceptable whether due to OCR or human intervention (such as rekeying text). Without professional editorial control, which will take the cost of digitisation out of most reasonable cost benefit ratios, there is a risk of occasional inaccuracy. In the case of a journal containing Nobel laureate papers, this risk may be considered unacceptable to the rights holder. There are several formats that allow just the image to be viewed or printed and also formats such as those used by JSTOR or 'image and text' PDF which allow the image to be viewed while OCR'd text is contained in hidden form 'behind' the image to facilitate searching.

One of the growing areas of serials digitisation is in support of electronic reserve collections for libraries. HEDS has experience with digitisation in relation to electronic reserves and the basic problem encountered is one of variance. I have already touched on rights, such as the authors moral rights and copyright. The difficulty for electronic reserves collections is the need to clear each item and the wide range of publishers involved across a whole collection. Most people involved in electronic reserves would like someone else to take all the stress away from copyright clearance. The issue for digitisation is very similar in that each article in the collection may be different in terms of layout, paper, print quality, language, use of diacritics or scientific notation and graphical content. The most efficient way to digitise is to change equipment settings or processes as little as possible over as large a set of documents as possible. However, it will be necessary to re-calibrate the equipment slightly as each article in an electronic reserve is scanned, to ensure that the end product has as little variance from the standard as possible. This adds a margin to the time and effort required to produce electronic reserve materials over other bulk processes. I support the idea of central provision of digitisation and copyright clearance services because in cases like this only centralised experience and resources can cope with the variance and range required to be cost effective and to meet high standards.

HEDS: futures and key benefits

In a short time HEDS has built up a wide ranging service and gained experience which will enable it to meet future clients' requirements. The main barrier facing most digitisation projects is cost. Digitisation has for too long been associated either with high costs or low grade end products. HEDS has the benefit of its project funding to ensure services are offered to HEIs at well below commercial rates. HEDS is also striving to provide a high grade service that matches the detail and standards of the end product to the vision of the clients requirements.

In conclusion, HEDS believes that the key benefits that its services offer are:

- the single point of contact for access to a range of digitisation services and management of the total package;
- the flexibility to tailor conversion advice, specification and delivery to suit the job in hand;
- competitive rates including heavily subsidised rates for HEIs during the project period;

- expert advice, consultancy and project management backed up by timely, reliable, quality conversion.

References

1. *Transactions of the Institute of British Geographers* samples. http://ppt.geog.qmw.ac.uk/tibg/tibg_forth/archive.html
2. Brailsford, Hugo. *Parallel Publishing for Transactions.* Ariadne. http://www.ariadne.ac.uk/issue11/ppt/
3. JSTOR Web site. http://www.jstor.org/aboutdemo/index.html
4. Sheffield University National Fairground Archive Web site. *World's Fair* newspaper. http://www.shef.ac.uk/uni/projects/nfa/newspap/wf/