THE HIGHER EDUCATION DIGITISATION SERVICE (HEDS): ACCESS IN THE FUTURE, PRESERVING THE PAST

Simon Tanner and Brian Robinson

Paper presented at the UKSG 21st Annual Conference, Exeter, April 1998

The aim of this paper is to discuss the following aspects of digitisation: HEDS and its services, with focus on how HEDS fits into the overall digital libraries initiatives of the UK; digitisation projects and resources in Europe and elsewhere; myths surrounding digitisation; and the future of digitisation in the UK. Many of the themes of this talk begin with 'it depends'. For example: it depends on what you want from the information content of the originals; it depends on the balance between costs and benefit goals; it depends on the nature of the original material itself.

Simon Tanner, Digitisation Consultant, HEDS, University of Hertfordshire and Brian Robinson, Manager, HEDS, University of Hertfordshire
E-mail: s.g.tanner@herts.ac.uk

Introduction to HEDS Services

HEDS has established 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 was established by the University of Hertfordshire with project funding from the Electronic Libraries Programme (eLib). This project funding was initially granted for two years from September 1996. From August 1998 HEDS will become a full JISC service and this will enable HEDS to develop existing services and expertise.

HEDS was established and funded because of a definite requirement within the UK higher education community for digitisation services. Difficult, high value projects and a larger scale conversion infrastructure could only be effectively achieved by a centrally resourced, national service. HEDS has already worked with the majority of the current eLib projects and with some public libraries, archives and museums, either providing technical advice or offering conversion 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 core services identified below:

- 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
- management of the complete job, from problem definition to final product delivery and acceptance
- prepararation of the functional and technical conversion specifications for the digitisation work

- provision of quality assurance procedures to validate the end product
- delivery of digitised materials on time and to agreed standards
- advice and assistance with copyright clearance
- raising awareness of digitisation for the higher education sector.

Digitisation projects

There are many exciting digitisation initiatives currently underway. The following projects are not intended to be a comprehensive list but they do include those that either HEDS are involved with or where there is a serials orientation.

The Parallel Publishing for Transactions (PPT) eLib project. HEDS are providing, in Adobe Acrobat PDF format, the complete backrun from 1935 onwards for the Transactions of the Institute of British Geographers[1] This resource will be made freely available to the UK HE community.

World's Fair Newspaper feasibility study[2] HEDS is currently completing a consultancy on the options available for digitisation. The newspaper is the main trade organ of the fairground community and is held as part of the National Fairground Archive at the University of Sheffield. The options explored have included scanning from the microfilm copies.

British Library of Political and Economic Science Pamphlet Collection. A large collection of political, social science and transport related pamphlets have been scanned and will soon be available from the London School of Economics. This project includes scanning items from tightly bound volumes and also from microfilm, and forms a unique information resource from an extensive collection.

Nature, one of the most prestigious of scientific journals, published by Macmillan Publishers Ltd, has technical piloting underway to establish the feasibility of digitisation of the back run. The project is being funded by JISC with Manchester Computing and HEDS in partnership to create and manage the digitised version. This will follow the general model established by JSTOR with images of each page viewed on screen with searchable text available. Manchester Computing have been at the forefront of developing metadata standards for the *Nature* digitisation that will have exciting implications for other journal metadata development.

JSTOR electronic journals. MIDAS at Manchester University have established a UK mirror site for the JSTOR electronic journal collection.[3] MIDAS are maintaining the JSTOR UK mirror site for UK higher education on behalf of JISC and in collaboration with the University of London. This improves access to a unique digital archive of core scholarly journals.

The Internet Library of Early Journals (ILEJ) is an eLib project aiming to digitise substantial runs of 18th and 19th century journals and make these images available on the Internet[4] The project has achieved good progress especially with popular journals such as the *Gentleman's Magazine* and *Notes and Queries.* The project is due to complete in August 1998, but hopefully will find further funding to continue the service.

DIEPER. - The Digitised European Periodicals project, is being funded by the European Community to address the need for a central access point where all digitised periodicals might be recorded.[5] Records of the register will be linked to reliable archives of periodical literature at different sites throughout Europe. In addition a search engine will allow for full text searching across the articles or at least the contents pages. The project partners are from all across Europe and whilst there is no formal UK partner, HEDS is associated with the project by providing supporting information as required. One of the participants is the Retrospective Digitisation Centre based in Germany who perform a similar service to HEDS for the HE community in Germany and are working especially on mathematics texts.[6]

The French National Library. The Bibliothéque nationale de France have made tremendous progress in getting a huge amount of French books, periodicals and pictures into digital formats for access in the National Library.[7] Working since 1992, they have converted over 27 million pages of text and 100,000 pictures into electronic formats. They have achieved this using

basic conversion standards and by converting large amounts of materials from microfilm stock. In 1997 a deal was struck with the publishers of the materials to allow access within the National Library and negotiations continue for wider access.

As can be seen from these short descriptions, there are a lot of interesting and influential digitisation projects underway at present, with a very diverse range of original materials being converted and made available. All these projects are using diverse techniques, data formats and standards for their digitisation. So, why is there is much diversity? Surely, such an IT orientated process can be narrowed to more quantifiable standards and more absolute methods.

I will now move to the myths of digitisation and explain why the answer is often; 'it depends' on the originals, your goals, budgets and cost benefits, rather than strict technical edicts.

Four myths of digitisation

> *"Why sometimes I've believed as many as six impossible things before breakfast." (Lewis Carroll - Alice's Adventures in Wonderland)*

The first myth to explode is that digitisation is a fully automated process. The idea that one places lots of documents into a document feeder, presses a button and out of the machine comes fully scanned, character recognised, indexed, tagged electronic files is sadly incorrect. There is a lot of human intervention at every stage of digitisation and this is where most of the money goes. Take, for example, the preparation of a 25,000 article short loan collection of photocopies. These articles might all be A4 single sheets in excellent condition - the ideal media for digitisation. However, there are firstly 25,000 staples to be removed at a cost of, say, 1p per document. This means that £250 has been spent to achieve nothing more than a pile of paper ready for scanning. Some commercial bureau will charge as much as 3-5p per document for preparation alone.

Therefore, a lot of thought must be dedicated to every stage in the digitisation process to ensure that the most efficient methods and most skilled operations are used. We must be conscious of the whole process, and, of course, the process will depend on your originals!

> *"Four legs good, two legs bad." (George Orwell - Animal Farm)*

There is an ideology in digitisation that is often misrepresented as the 'more is better' approach to all scanning. This is especially true of image resolution. Resolution is the description given to the number of dots or pixels per inch that represent the original in an scanned image file. Increasing the number of pixels used will result in higher resolutions and a better ability to record fine detail. There is an archive standard of 600 dots per inch (DPI) for black and white text suggested by Cornell University and JSTOR for their projects. They are exactly right for the type of material and purpose to which they are referring. However, this is frequently misrepresented as the de facto standard for all types of materials, for all purposes. There is a point, though, at which adding resolution does not add detail or content that is value-added. So there are several reasons in my "it depends" theme that suggest not going to higher resolutions automatically:

- where file size for the target application is a major issue;
- where the purpose is to represent the content accurately rather than reproducing the original in every detail;
- where the content is to be repurposed for character recognition;
- where adding detail or definition does not actually add content or value;
- where the original material does not require it, e.g. examination papers;
- where you cannot afford the cost of 600DPI.

We must guard against specifying requirements purely because we are following an edict; we must choose the most appropriate technology for our needs.

> *"In all pointed sentences, some degree of accuracy must be sacrificed to conciseness." (Samuel Johnson - On the Bravery of the English Common Soldier 1797)*

Almost nothing else in digitisation causes as much debate as the usefulness and accuracy of optical character recognition (OCR). On the one side are the OCR converts who believe that everything is suitable for OCR, including 18th century or middle

English texts and at the other extreme are the deeply sceptical. The answer is, as Samuel Johnson suggested in a different context, if you want to do OCR quickly and cheaply then prepare to sacrifice accuracy. If OCR text is being repurposed as the means of viewing the content of the electronic file, it must be corrected or at least checked by humans. The cost of correction grows in relation to the condition of the original and the way the text has been printed. However, OCR to gain text to feed an index to enable searching usually requires no correction. This is because the OCR accuracy does not have to be so high to enable good results with a fuzzy search engine.

> *"Pickering: Have you no morals, man?*
> *Doolittle: Can't afford them, Governor."*
> *(Act II, Pygmalion, George Bernard Shaw)*

I have probably left you with the impression that digitisation is a very expensive process. However, just like Elisa's father we tend to view things as expensive only as long we do not value them. Digitisation has for too long been associated either with high costs or low grade end products. If the product is not highly valued then digitisation is bound to seem expensive. When one considers the actual value added to the information resource in terms of wider access and preservation, it has to be balanced against the costs for the very large amount of work done to convert the material. Costs are also dropping as the technology improves and this opens the door to projects that would not have been considered previously,. In additon, standards of product are rising all the time. So digitisation is a decision that should be made with an eagle eye on the value and benefits gained versus the money to be expended.

The future of digitisation in the UK

In conclusion, I would like to address a number of issues that I think will be important to ensure the continued development of digital library resources in the UK.

- *Continued investment in the conversion infrastructure.* Resources are scarce and funds will be best spent on developing centralised resources or using vendors, rather than every project purchasing equipment with limited lifetimes and funds.
- *Technology catch up.* The equipment and software, needed to achieve the high standards of conversion required to serve the needs of the Higher Education communities, are just now becoming available and ever more efficient with less and less human intervention required. This will drive down the cost of conversion whilst increasing the standards of images. I hope that this will encourage further projects to digitise materials previously inaccessible to conversion.
- *Maximise value of materials.* There is a trend towards high profile, large scale conversion projects that will benefit all of Higher Education in the UK. These are extremely important, but I hope that the funding bodies do not remove money from the smaller projects with a very focused user group in mind, where the costs are low, but the value added very high.
- *Planned growth of digital libraries.* eLib has developed many resources to create digital libraries but much of this has come about through the strong interests and work of a few people without institutional review. There is a need to resolve digitisation priorities, based on experts within disciplines determining the core resources that would be most valuable for teaching.
- *Scalability.* Again eLib has been successful in creating small scale digital libraries. However, as the digital resources grow we must plan to ensure that there is sufficient capacity to manage and maintain large sites and large electronic collections.
- *Access issues and metadata.* The future costs for conversion will continue to drop, but the costs of creating intellectual access and metadata will remain high. It is important to develop appropriate levels of content description to enable access mechanisms, if the value of the digital library resources are to be fully realised.

There are many wonderful information resources that make up the treasure house that is the library and archive system. It is likely that many of them will be digitised and there is certainly a trend towards treating digitisation as a serious prospect. As evidence I point to the £30

million over five years expenditure on digitising special collections recommended in *New library: the people's network* recently published by the Library and Information Commission.[8] It is also interesting to note that *Virtually new - creating the digital library: a review of digitisation projects in local authority libraries and archives* recently recommended the establishment of an "agency to advise and co-ordinate public library digitisation" and that there "must be substantial new, targeted external funding if any significant volume of work is to proceed".[9]

HEDS: Futures

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 remains money. 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 and will be very competitive for non-HE clients. 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, I hope I have out across the main message that digitisation is not solely about technology, but more importantly, it is about your information goals and needs. It is vital to have a clear idea idea of what you want to achieve from your resources, to relaise that you are limited only by the nature of your originals, and to understand the cost benefit ratios of the conversion to electronic format. Then call HEDS and we will see what we can do to help!

References

1. Parallel Publishing for Transactions (PPT) http://ppt.geog.qmw.ac.uk/tibg/tibg_forth/archive.html
2. World's Fair Newspaper feasibility study http://www.shef.ac.uk/uni/projects/nfa/newspap/wf/
3. JSTOR http://www.jstor.ac.uk
4. Internet Library of Early Journals (ILEJ) http://www.bodley.ox.ac.uk
5. DIEPER - the Digitised European Periodicals Project http://www2.echo.lu/libraries/en/projects/dieper.html
6. The Retrospective Digitisation Centre, Germany http://www.sub.uni-goettingen.de/GDZ/
7. French National Library http://www.gallica.bnf.fr/
8. *New library: the people's network.* http://www.ukoln.ac.uk/services/lic/newlibrary
9. *Virtually new: creating the digital collection.* http://www.ukoln.ac.uk/services/lic/digitisation/intro.html