

Digitisation: How much does it really cost?

Simon Tanner and Joanne Lomax Smith, HEDS.

Paper for the [Digital Resources for the Humanities 1999](#) Conference held at King's College London, 12-15 September 1999.

Introduction

This paper explores the most thorny issue of digitisation - how much it costs.

Digitisation is sometimes perceived as being an expensive process. Possibly because it means buying equipment and utilising staff and space when working in-house or maybe because outsourcing means finding real and visible funds that disappear from your budget in a rather large lump. Other comparisons made are that you only outsource for high volume, low value materials like exam papers and that in-house is the best way to achieve high quality at reasonable costs. However, the reality as opposed to these perceptions is not so clear cut.

The Higher Education Digitisation Service (HEDS) has a unique viewpoint regarding digitisation costs. We exist both to advise our clients on how to optimise their in-house resources for a digitisation project and also to act as a production service for clients who wish to outsource the digitisation process. The sum total of the real costs for either approach can therefore be explored from our experience.

We look at the following in this paper:

- What makes up a digitisation project and what are the major cost factors?
- How will factors such as the choice of original and the technical specification will effect costs?
- Some reasons for choosing whether to outsource or use in-house resources.
- Examples of comparative costs between the two approaches.

When referring to digitisation in this paper then we mean the process of scanning to gain a data image. The paper will not deal with optical character recognition, mark-up, SGML, watermarking or any of the many post scanning digitisation techniques as this would require several hours/pages, rather than the 30 minutes available for this paper.

HEDS are not going to provide a huge list of various materials and unit prices for digitisation. This is just not possible in the real world. The reasons for this will become increasingly apparent as this paper progresses, but a swift explanation of why pricing cannot so easily be bracketed is required now. Right from HEDS conception in 1996, we have established an "it depends" answer to this question. There are 3 threads:

- It depends on what you want from the information content of the originals.
- It depends on the balance between costs, technology and benefit goals.
- Most of all - it depends on the nature of the original material itself.

However there are some pricing examples available:

HEDS recommends Dr Stuart Lee's paper, which includes real-life project costs (including some HEDS projects) and comparisons on in-house versus outsource costs:

<http://www.bodley.ox.ac.uk/scoping/digitization.html>

This report by Tony Hendley also has indicative pricing models for content creation:

Hendley, T. Comparison of methods and costs of digital preservation. JISC/NPO Studies on the Preservation of Electronic Materials. British Library Research and Innovation Report, 106. London: British Library Research and Innovation Centre, 1998.

What HEDS will offer at the end of this paper is a costing matrix. This will identify for various types of materials the cost factors and rank them relative to each other. In this way it will be possible to assess whether A4 exam papers will take more or less resources (either in-house or outsourced) than something more tricky such as glass plates.

We begin by looking at the Digitisation Project itself:

The Digitisation Project

The basic tasks in most digitisation projects are similar wherever the actual digital conversion takes place. Broadly speaking these are:

1. Assessment of the need for digitisation.
2. Selection of materials.
3. Deciding what you want to achieve from the information content of the originals.
4. Deciding how you are going to reach the end product.
5. Finding the funds for the project.
6. Planning the project and assigning resources.
7. Preparing the originals for digitisation.
8. Conversion.
9. Quality assurance checks to ensure the output conforms to specification.
10. Return originals to their place in the collection.
11. Mount data.
12. Make provision for archiving and preserving the data.

Within the overall project framework the actual conversion work is a relatively minor part of a more complex, person-driven scheme. Here we identify the first and most important cost factor - people. It is obvious, but the more person hours required to complete any of these tasks the higher cost is likely to be. As many aspects of digitisation may potentially be automated some of the conversion costs may be mitigated - but almost none of the other project aspects can be made any cheaper because they are people intensive. However, the staff costs are often hidden, absorbed by the organisation and rarely fully funded by the project budget. Cost cannot always be defined purely as the project budget bottom line - the figure put into a funding bid may not cover the full cost of the time and effort that goes into a digitisation project.

You may also have noticed the order of the tasks identified here and noted that finding funds was fifth on the list. One of the main reasons for digitisation appearing expensive is that so many projects start with funding rather than with the originals and the goals for the information content. Thus funding is often assigned on a finger in the air basis and when the resources come to be used the conversion part looks large in comparison to the available budgets and appears expensive.

Another problem can be poor estimating - usually based on inexperience or the assumption that the process is very simple. HEDS was asked a couple of years ago to go speak with a Chief Executive who had ordered the Librarian to get all the 80,000 organisation reports digitised. He had rung a bureau and asked how much for 80,000 documents - to which they replied (hearing 80,000 pages, not documents) that it would cost about £8,000 for A4, black and white at 200 dpi resolution (roughly 10 pence per page side). In actuality, there were 8 million pages (each report had about 100 pages) and that Chief Executive had a shock when informed it would cost more realistically in the region of £800,000.

The Basic Costs

Here we also have an indication of some basic costs - those costs that are going to be incurred by every project. These are staff costs in terms of project planning, preparation of materials, some level quality assurance and checking of output and the return of originals to their place in the collection.

By far the greatest of these costs is preparation - this might include:

- The movement of materials from one place to another - requiring inventories and packaging for movement (even when working in-house this is recommended).
- The time taken to assign unique identifiers to originals if this hasn't already been done.
- The cost of removing staples, or cleaning transparencies or otherwise preparing the physical items.
- The cost of clearing copyright or other rights to use materials.

MR Data Ltd - a commercial bureau - have estimated that preparation is accounting for up to 30% of total project costs for some of their complex projects.

When considering outsourcing you might want to look at the preparation requirements and rather than just handing over a huge pile of originals, look at what you can do in-house to reduce the cost at the supplier end of things. Alternatively, you might find that the outsource agency has far cheaper staff costs and can do a lot of preparation work such as de-stapling and disbinding far cheaper than you can. The measure here is to only outsource that which does not require great initiative or skill as this tends to cost more - if you cannot easily write a specification or set of rules for how you want it done then expect to pay more for someone else working it out for you!

HEDS recently looked at a set of university student records that the university in question hoped to have in digital form for alumni record purposes. The records for the early part of this century were in a dank basement stuffed in several plastic laundry baskets. They were in no particular order, poor condition and had no unique identifiers such as record or student number. Therefore they would have to be sorted, cleaned and identified in some way. The basic cost of scanning each card was approximately 11 pence per item, but the overall cost rose to 28 pence per card and the indexing costs were a further 8 pence. What looked initially like an attractive prospect was deemed too expensive purely due to preparation costs.

The Technical Costs

When we focus narrowly on the pure costs of scanning we can say that there are two cost elements:

- The cost of handling or otherwise moving the original material through the scanning process.
- The cost of writing an output file to the required resolution, bit depth and quality.

The cost of handling is mainly related to the amount of automation that is possible in a process. Some examples:

- A4 laser printed sheets can be passed automatically through a scanner using sheet feeders.
- A bound volume will need every page individually turned on a bookscanner.
- Photographs cannot be automatically fed through a scanner because they will be damaged and jam the mechanism.
- 35mm mounted slides may potentially be loaded into a carousel for automated scanning.
- Glass plates can take up to 15 minutes each to get the plate onto the scanner, take a scan and then remove the plate back to its holder.

Because increased handling means more human intervention then costs will rise. Automation of the transit of materials through a scanner will reduce unit costs. However, it should be noted that prices of 7-10 pence per A4 page or 20 pence per microfilm frame scanned are based on bureau/production facilities with very expensive machinery to enable large volume automated scanning.

The drum scanners used commercially to produce up to 600 dpi black and white scans have scan speeds of less than a 3 seconds per sheet with autofeed for several hundred sheets and thus may cost over £25,000. Microfilm scanners start at £60,000 for basic automated versions with prices rapidly going through the £100,000 mark for full automation and greyscale capability.

So whilst an in-house operation might be able to afford a robust scanner with a 50 sheet feed mechanism for less than a thousand pounds the levels of throughput are not comparable and thus costs will differ due to scan times being more like 10-15 seconds per scan and the sheet feed needing more frequent reloading.

The output specification will have a distinct impression upon price. The core issues here are resolution, bit depth and quality.

Resolution: this is the amount of detail included in the output image measured in dots per inch (dpi). Depending on your original a choice of a suitable resolution is required so as to gain the appropriate level of information content. For good quality printed A4 paper a low resolution of 200 dpi might be suitable to merely read the content - a standard used in business applications. However Cornell, JSTOR and TASI would all recommend 600 dpi to gain a more forensic accuracy. 600 dpi costs a lot more than 200 dpi - so for exam paper type projects 200-300 dpi would be cost-effective, but for rare books then 400-600 dpi would be more sensible as the output would match the required end-use.

More resolution means:

- slower scan times.
- more time to write data because the file sizes are bigger.
- more expensive equipment.
- more delivery media such as CDROMs.
- more overall human time in managing the operations.

Doubling your resolution means much more than doubled file size. So for an A4 sheet in Black and White uncompressed TIFF:

- 150 dpi = 250 kb
- 300 dpi = 1 Mb
- 600 dpi = 4 Mb

Prices may in many specifications start doubling as the resolution increases - roughly from 7 pence per page to 15 pence per page to 28 pence per page for the above simple specification.

Bit Depth: this is the amount of information per pixel and also relates to colour information.

- 1-bit is generally black and white.
- 8-bit is generally for 256 greyscales although it can be used for low quality colour.
- 24-bit is generally used for high colour images.

The issues here are exactly the same as for resolution - the higher the bit-depth the more information content the higher the file size and higher the price. However, the price increase from increasing the bit depth is separate from any incurred with resolution increases and therefore it is an additional cost factor.

So for an A4 sheet uncompressed TIFF at 300 dpi the file sizes go:

- 1-bit (B&W) = 1 Mb
- 8-bit (greyscale) = 8 Mb
- 24-bit (RGB colour) = 24 Mb

These file sizes are pretty big and will have a consequent impact on price. However if we increase the resolution from 300 to 600 dpi as well then going from 1-bit to greyscale the file size jumps from 1 Mb for B&W 300 dpi to 32 Mb for 600 dpi greyscale.

As you can imagine there are thousands of combinations of resolution and bit-depth alone, even completely ignoring the technical treatments, the handling and image quality issues. It may now become clearer why HEDS do not produce standard price lists - it is important to price each project based on the goals of that project. However similar to other projects your work may be your project goals will always be unique and thus so should your specification.

Reasons to choose to outsource or use in-house resources.

There are many reasons for sending materials to an outside agency for digitisation. Cost is a major one - if you have very large volumes outsourcing will generally be cheaper than setting up in-house processes. But there are other reasons such as that the originals are not capable of being scanned successfully in-house - such as for bound volumes which require a book cradle. Sometimes the project may be just too complex to be completed with the available in-house resources - such as projects requiring advanced colour management skills for colour artefacts.

The project manager may decide to use in-house resources for several reasons:

- The collection cannot be moved out of the institution.
- The collection is badly organised and needs skilled reorganisation as an integral part of the process.

- The digitisation needs to be phased in small amounts over a long period.
- The digitisation tasks and goals are very simple.
- Where the volume of work is very small.

There are some base line infrastructure requirements to in-house digitisation:

- A robust production level scanner.
- A powerful PC with lots of memory (at least 256Mb RAM).
- Plenty of system resources such as backup and write to media (e.g. CDROM) capacity.
- Software to assist the digitisation.
- Experienced staff to run the equipment and staff to oversee the process.

The scanner and PC may be used year round by other projects so investment in them may not be just for a single project. Money will have to set aside for the regular maintenance and upgrade of all equipment and software. The member of staff will need full training in order to run the scanning procedures and if casual staff are used this training may need to be repeated for each of them. Supervision will be required by a member of professional staff and the data output will require extensive quality checking as well. HEDS estimates that the start up capital outlay for a fully fledged digitisation workstation with basic production standard flatbed scanner and software aimed at scanning paper will be about £3-4,000 and that to cope with photographs or colour would require about £6-8,000 in initial outlay. This is assuming that the in-house operation wants to approach anywhere near the unit price of production available from outside agencies such as HEDS.

As promised earlier in this paper HEDS have constructed a costing matrix. This identifies for various types of materials the cost factors and rank them relative to each other. In this way you will be able to make your own relative assessments whether considering an in-house or out-sourced process.

The HEDS Matrix of Potential Cost Factors

Below is the HEDS Matrix. There are some assumptions inherent in this matrix, the first being that all the original materials are in excellent physical condition and have no specific problems in terms of handling or scanning other than their intrinsic physical nature. The condition of your originals has a marked effect on what can be done and so this should be included in any factored cost analysis gained from using the HEDS Matrix.

HEDS Matrix	Materials							
Cost Factors	Printed A4 Paper (B&W)	Bound A4 Volumes (B&W)	Printed A4 Paper (Colour)	35mm Microfilm (B&W)	Photo prints 5"x4" (Colour)	35mm slides (Colour)	Negative photo film unmounted (B&W)	Glass plates 5"x4" (B&W)
Typical Spec: Resolution & Bit/Colour Depth	300 dpi 1-bit B&W	400 dpi 1-bit B&W	300 dpi 8-bit colour	400 dpi 8-bit greyscale	600 dpi 24-bit colour	2700 dpi 24-bit colour	2700 dpi 8-bit greyscale	600 dpi 8-bit greyscale
Preparation	LOW	HIGH	LOW	LOW	MEDIUM	MEDIUM	HIGH	MEDIUM
Handling	LOW	HIGH	LOW	MEDIUM	HIGH	LOW	HIGH	VERY HIGH
Automated Scan	✓	✗	✓	✓	✗	✓ ✗	✗	✗
Operator Skills	LOW	MEDIUM	LOW	MEDIUM	HIGH	HIGH	HIGH	MEDIUM
Post Processing Costs	LOW	LOW	MEDIUM	MEDIUM	HIGH	HIGH	MEDIUM	MEDIUM
Resource Costs	LOW	MEDIUM	MEDIUM	VERY HIGH	MEDIUM / HIGH	MEDIUM / HIGH	MEDIUM / HIGH	MEDIUM
QA costs	LOW	LOW	LOW	MEDIUM	HIGH	HIGH	HIGH	MEDIUM
Filesizes	LOW	LOW / MEDIUM	MEDIUM	MEDIUM	HIGH	HIGH	MEDIUM	MEDIUM
Overall	LOWER	MEDIUM	MEDIUM	LOWER / MEDIUM	HIGHER	MEDIUM / HIGHER	HIGHER	VERY HIGH

Overall Ratings: Lower ~ £0.10 - £0.20 per unit item
 Medium ~ £0.20 - £1.50 per unit item
 Higher ~ £1.50 upwards per unit item

Note the anomalies in this matrix:

- The glass plates would be expected to be rated in the medium bracket, but the handling is so high that this skews the whole assessment into the higher cost bracket. This sort of anomaly can happen for any original material and should always be kept in mind.

- That the rating for microfilm is low/medium when the resource costs and very high. This is based on the assumption of a high volume of throughput. The costs of scanning are relatively low, but the cost of equipment is very high. For any operation, whether in-house or outsourced, the volume of microfilm expected to be put through the scanner would have to be very high (many tens of thousands per year) to justify/recoup the cost of purchase.
- The differences in photographic film transparency format has interesting affects on rating, due mainly to their handling characteristics and physical size. Broadly speaking the mounted photographic transparency will be easier, quicker and cheaper to process than non-mounted items.

Conclusions

There are many aspects to take into account when estimating the pricing for any digitisation project. As has been described in this paper there are many combinations of resolution and bit-depth alone, even completely ignoring the other technical treatments, the handling and image quality issues. HEDS hopes it is now clear why standard price lists are often misleading and why it is vital to price each project based on the goals for that project. However similar to other projects the work may appear to be, everyone's project goals will always remain unique and thus so should the specification and cost estimation.

This is where HEDS comes in as a friendly source of guidance and independent advice based on our extensive experience. We are available to provide free initial advice and support, plus a charged consultancy service for more detailed work. We may be contact via email, phone or our Web page (listed below).

We welcome any feedback upon the HEDS Matrix and are looking to extend and improve the content as time and project deadlines allow.

We may be contacted at:

Email: HEDS@herts.ac.uk
Phone: 01707 286078
Web Address: <http://heds.herts.ac.uk>

HEDS
University of Hertfordshire
College Lane
Hatfield
Hertfordshire
AL10 9AB