Handbook on Cost Reduction in Digitisation

September 2006

Handbook on Cost Reduction in Digitisation

September 2006

General co-ordination
Rossella Caffo (MINERVA and MINERVA Plus Project Manager)
Antonella Fresa (MINERVA and MINERVA Plus Technical Coordinator)
Pier Giacomo Sola (MINERVA and MINERVA Plus Organisation Manager)

Author
Simon Tanner
(King's College London, http://www.digitalconsultancy.net)

Editorial Committee
Antonella Fresa (Italy)
Borje Justrell (Sweden)
Dov Winer (Israel)

Contributors
Alojz Androvic
David Dawson
Kate Fernie
Anne Grady
Lennart Hedlund
Dimitrios Koutsomitropoulos
Anton Pärn
Marzia Piccininno
Orly Simon
Maria Slivinska
Pier Giacomo Sola
Joan Ward

Secretariat
Marzia Piccininno

Web version and editing
Maria Teresa Natale and Andrea Tempera
http://www.minervaeurope.org/publications/costreduction.htm

2006 MINERVA Plus Project

Table of Contents

Foreword 5

1 Introduction 7

2 Overview 9
2.1 Methods to reduce costs 11

3 Reducing the Cost of the Workforce 13
3.1 The cost of human intervention 13
3.2 Reducing human intervention 13
3.3 Reducing the cost of labour 14
3.4 Outsourcing 16
3.5 Case study: The British Library 18

4 Automation 19
4.1 Mechanistic automation 19
4.2 Software based automation 21
4.3 Case study: Automating metadata capture 24

5 Selection and Preparation for Digitisation 27

6 Increase Performance and throughput 29
6.1. Workflow 29
6.2. Case study: National Archives of Sweden 30

7 Continuous Improvement and Quality Assurance 33
8 Conclusion 35

About the author 36

Foreword

This *Handbook on Cost Reduction in Digitisation* has been conceived in the frame of the activities carried out by the working group on best practices of the MINERVA network.

Some general objectives were associated with the working group, i.e. to support the development of skills and to increase efficiency encouraging take-up of good practice and promotion of "centres of competence". A first outcome of the working group was the *Good Practice Handbook* published by the MINERVA project in 2004, distributed in printing and online all over Europe, translated in several languages and considered the “best seller” of MINERVA.

The successor of MINERVA, the MINERVA Plus project, identified the cost reduction as a specific “hot” topic to be investigated through the analysis of the work flow and production chain of digitisation.

A workshop took place in Fransta in January 2005, in the middle of Sweden, where the National Archives of Sweden has a scanning facility, the Media Conversion Centre (MKC) with around 80 employees. MKC scan hardback books and loose sheets in formats from stamps up to A0 in b/w, greyscale and colour. The scanning production at MKC is for the moment (2006) 90 000 images per day. About 500.000 images are scanned each year as 1-bit 600 dpi files in A4 format using automatic feed scanners. The cost for each scanned file is 0.10 Euro. 1/3 of the cost goes to scanning (data capture) while preparation, image control, extras and administration are the four other almost equal major activities.

Starting from a very practical analysis of the work of MKC, the idea was born to produce a handy instrument to be given in the hands of the people who are actually planning and working on digitisation programmes and projects, with the aim to improve their productivity and to contribute in such a way to accelerate the achievement of a critical mass of digital content, that is in fact the most important step towards the Digital Library.

Digitisation projects are easy to plan and fun to conceive. The possibilities seem to be endless and the only limitation is the sky. The

tremendously huge number of ways to use information in digital format may appear both exciting and cheap. The excitement about what may be achieved sometimes tends to over shadow the facts of costs, not necessarily the costs connected to the digitising process itself, but the costs for sustainability, in other words to keep the created images files alive so they are accessible and usable over time. A digitisation project without proper planning may turn into something like a black hole in the sky.

Antonella Fresa
Borje Justrell
Dov Winer

1 Introduction

The effective utilization of resource is amongst the most important management activity in developing digital content and establishing digital collections. History has many examples of great cultural collections being created through the largess of benefactors, or the tax paying public, without necessarily having much consideration to cost effectiveness. Between 1881 and 1917, Andrew Carnegie contributed $56m to build 2,509 libraries. In modern times there are new market economies facing today's manager, and competition for attention and resources because of initiatives such as Google's BookSearch. This means that every last drop of value must be squeezed from the resources available in order to maintain funding now and in the future.

The effective utilisation of resources in digitising collections, developing digital content and establishing digital resources equals:

- the start up costs of creating or purchasing digital content;
- the implementation costs for establishing access; and
- the implicit costs in managing and maintaining digital resources into the future.

For digitisation, the key factors influencing cost are:

- the nature of the original item to be digitised;
- the digitisation processes and mechanisms possible; and
- the information, content and delivery objectives to be achieved.

This guide is a practical instrument for cultural institutions to support decision making and planning for digitisation initiatives. It covers the key parameters to be taken into account to assess and monitor the cost of digitisation, how to model the cost of a given initiative, and ways in which the cost of digitisation can be reduced.

.

2 Overview

When planning the costs for a digitisation project there are many potential elements. This guide will focus upon the main components to be considered, including:

- **Selection**: choosing what will be digitised.
- **Preparation**: making objects ready to be digitised.
- **Metadata creation**. Examples:
  - Cataloguing
  - Description
  - Indexing
  - Administration and technical information.
- **Preservation/conservation** of the physical object.
- **Production of intermediates**. Examples:
  - Microfilming
  - Photography.
- **Technical infrastructure**. Examples:
  - Network for moving data between production and quality checking machines.
  - Data storage, backup and transfer.
  - Digitisation infrastructure, such scanners, computers, suitable physical space for equipment and activity and software.
- **Conversion to master digital format**. Examples:
  - Scanning
  - Direct digital capture, e.g. digital camera
  - Audio or video encoding.

- **Production of surrogates for end use**. Examples:
    - Optical Character Recognition or rekeying for full text searching
    - Compression and resizing of images for Web use, e.g. thumbnails
    - Video and audio compression for Web use, e.g. streaming or mp3 download.
    - XML or other mark-up to content for Web use.
- **Quality control** and assurance for images, text and metadata.

There are several guides and case studies to calculating costs and defining the digitisation cost elements which provide a good starting point. Many of these are compiled or referenced in the NINCH resources list for the Price of Digitization Symposium in 2003[1].

The RLG also provide an extremely useful worksheet that helps to identify digitisation cost factors across the following elements and steps[2]:

1) Select materials

2) Determine the size of the collection

3) Prepare documents

4) Determine imaging requirements (benchmarking)

5) Determine requirements for and create metadata

6) Determine imaging costs

7) Determine text conversion costs

8) Determine SGML encoding costs

9) Determine Finding Aid Conversion Costs

10) Post-process digital files

11) Estimate additional local costs.

[1] *The Price of Digitization: Resources - NINCH SYMPOSIUM: April 8, 2003, New York City*, <http://www.ninch.org/forum/price.resources.html>.
[2] *RLG Worksheet for Estimating Digital Reformatting Costs. A guide to the preparation of a budget for digitisation*, <http://www.rlg.org/en/pdfs/RLG Worksheet.pdf>.

2.1 Methods to reduce costs

In a business context, the normal methods used to reduce the costs associated with making a product are:

- Reduce the cost of the workforce;
- increase productivity and throughput;
- increase performance and efficiency; and
- add value by improving quality.

The most frequently used methods for reducing costs in digitisation are:

a) Reduce the cost of labour;

b) Automate to reduce levels of human intervention in digital conversion and metadata creation;

c) Select and prepare originals to enable higher volumes and reduce variation in the workflow;

d) Increase overall performance and throughput to make the most efficient use of capital expenditure; and

e) Continuous improvement and optimisation through rigorous quality assurance.

3 Reducing the Cost of the Workforce

3.1 The cost of human intervention

The higher the level of human intervention in a digitisation process the greater the likely costs associated to that process.

For instance, the key cost difference between scanning glass plate photographs and 35mm plastic mounted slides is that glass plates take more staff time and skill to handle safely than the 35mm slides. They are both photographic materials, but the originals physical nature forces different modes of handling and scanning mechanism. Thus, glass plates take several multiples more staff time to image and, assuming everything else is equal, glass plates will be much more expensive to image than 35mm slides purely because of the additional human costs.

So the key methods for reducing costs must involve:

- reducing the level of human intervention; or
- lowering the total cost of that labour.

3.2 Reducing human intervention

One means of reducing the level of intervention may be achieved by automating the scanning mechanism or metadata creation process and thus reduce the time expended to a minimum. Automation is discussed in detail within Section 4.

Other cost reduction methods for lowering intervention are covered later in this guide: such as the selection and preparation for digitisation (Section 5) or increasing performance (Section 6).

Whatever the process, there may be a way to make it more efficient without compromising quality. Even assigning descriptive metadata or subject headings, a process that is often hard to automate and needs skilled human intervention, can be made quicker by using a controlled vocabulary and enabling drop down lists for data entry.

The rule for cost reduction is to look at every stage that requires human intervention and either remove it, reduce it or make it as efficient as possible.

3.3 Reducing the cost of labour

The true cost of labour is not just defined by the salary as there are many other costs issues to consider such as total number of productive hours, holiday entitlements, pension and benefits payments, equipment provided, space used, support required etc. A quick way to get an estimate of these total costs of labour is to multiply the basic salary by 170%.

As increasing productivity is covered in Section 6, this section will primarily focus upon the salary cost and outsourcing.

Reducing the salary cost

To reduce the salary cost usually means that the task is made easier so that lower skilled and less expensive staff are needed. Alternatively, it means seeking a cheaper workforce to carry out the tasks.

Deskilling the task can be achieved by a number of means:

Breaking the activity down into modules

The objective is to redesign the activity so that highly skilled or expert staff are only used when absolutely necessary. Then lower qualified (and thus lower paid) staff can carry out repetitive or straightforward tasks. This can work even in complex digitisation projects where the technology or task is intellectually challenging.

Example:

A large scale plant specimen digitisation project was employing post-graduate qualified plant taxonomists. These plant experts were scanning the specimens and then entering plant information (from written notes on the specimen mount) to a database. This is an example of highly qualified staff doing tasks that could be achieved by other lower paid staff. The image scanning can be separated out as a repetitive task that could be done in batches by trained but lower paid staff. The database entry does need plant knowledge, but even this could be trained to a well designed set of procedures with a few taxonomists available to advise on difficult cases. By breaking the activity into modules costs can be reduced by only using highly paid staff to do tasks that warrant their valuable time.

Provide better tools and guidance

Digitisation is often considered high cost because the technology appears so advanced that highly skilled staff is required to operate the equipment. Firstly, as shown above, the skilled staff can be used to support other staff rather than requiring everyone to be equally skilled. Secondly, the use of better tools and proper guidance and procedure manuals can significantly reduce the skills needed for many tasks.

Two examples:

1) The use of modern calorimeters, colour charts and colour profile management software means that digitisation equipment and environment can be calibrated to be colour accurate. The staff need only regularly maintain this calibration through the use of the colour tools provided thus reducing the skills needed for a task previously requiring expert knowledge to be continuously applied. See Gretag Macbeth[3] or the International Colour Consortium[4] for more information.

2) Drop down lists of acceptable terms for data entry or a tool that automatically adds the appropriate XML tags to data entered are another example of simple tools to reduce the skill level required to achieve a task. They also tend to increase the quality and consistency of the end product as well.

These examples illustrate that reducing the skill levels need not be expensive, difficult or patronising to staff. It just requires a consistent enough flow to the digitisation work so that procedures can be designed and tools optimised to make them as easy as possible to use.

Invest in training

This may seem to contradict the ethos of reducing skills to reduce costs! In fact, the suggestion is that a lower paid member of staff with appropriately focused training may be able to achieve the same performance in a narrowly defined activity as a more costly member of staff. Care should exercised that they are not expected to be as skilled in every aspect of the task as that other level of staff, otherwise they should be paid the same salary.

In many digitisation projects, much of the work is roughly 85% repetitive or routine, with 12-15% difficult and <1% very difficult. Rather than appoint staff who can deal with all these aspects, it is possible to appoint at a lower grade and train the staff to deal with the narrow set of issues that make up the 12-15% difficult cases. The remaining few problems can be resolved by experts either in house or from outside the project.

[3] <http://ww.gretagmacbeth.com>.
[4] <http://www.color.org>.

Investment in training can also improve the workflow and throughput whilst reducing the costs associated with rework from quality defects.

3.4 Outsourcing

Seeking a cheaper workforce is one key reason for using an outside agency for digitisation. For large volumes, outsourcing will generally be cheaper than setting up in-house digitisation processes.

Why not to outsource

There are some circumstances in which considering in-house digitisation will continue to offer advantages for practical and cost reasons, such as:

- the collection is difficult to move or cannot be moved outside of the institution;
- the collection is badly organised, not inventoried or un-catalogued to the item level and needs skilled reorganisation as an integral part of the process;
- the digitisation needs to be phased in relatively small amounts over a long period;
- the preservation handling of the originals cannot be satisfactorily achieved in the outsourced environment;
- the digitisation tasks and goals are very complex and varied; and/or
- the volume of work is very small.

Utilising an internal digitisation unit gives an institution the value of equipment, highly trained staff plus the movement and treatment of the originals can be closely controlled. It also avoids the tendering and procurement process which can be onerous and expensive in its own right.

A further reason why many projects are undertaken in-house is that the staff time, overheads and some consumables such as file storage can often be swallowed up by the institution and do not become apparent as a cost factor of the project, thus making this appear to be a cheaper option than outsourcing. In this scenario, it is only when outsourcing becomes an option that all the internal overhead costs become visible to the organisation.

How outsourcing can reduce costs

Using an external supplier means the equipment and expertise of a third party can be exploited, while the project team concentrates on their specialist area of the project.

Using a bureau also means that the cost of buying and maintaining specialist and expensive equipment is not fully borne by the project. Many projects never utilise the full depreciation value of the equipment they buy and this inflates the cost of the digitisation activity.

Other reasons for outsourcing might include:

- a large volume of work to be done in a short period of time;
- excessive cost of specialist equipment (such as bound volume or microfilm scanners);
- Lack of capability - unable to deliver the quality needed in-house due to lack of skills and experience;
- the project has space, infrastructure or staffing constraints which preclude in-house digitisation; or
- to take advantage of overseas bureau (with lower staff costs reflected in their prices) for activities such as bulk paper or microfilm scanning, text rekeying or XML mark-up.

Outsourcing may genuinely reduce digitisation costs and provide strategic advantage. To test whether it will do this, the following questions should be answered:

a) Will you be able to refocus on your core skills and use the vendor's expertise to reduce overall costs?

b) Will your competence be enhanced and your capacity to complete the project improved?

c) Do you have increased access to key technologies and equipment that would otherwise be unaffordable?

d) Have you reduced the risk and cost of obsolescence?

e) Have you enhanced economies of scale in terms of human resources?

f) Have you increased your access to skilled personnel?

g) Will outsourcing increase your control of expenditure, expenses and overheads?

3.5 Case Study: The British Library

The British Library have 2 major digitisation projects that identify the core issues for outsourcing digitisation. These projects are:

Archival Sound Recordings[5].

This seeks to deliver up to 12,000 segmented audio encodings totalling 3,900 hours of sound recordings from distinct and unique collections.

19th-Century British Newspapers[6].

Up to 2 million pages, totalling approximately 10 billion words of British newspapers from 1800–1900.

Both of these projects have outsourced the digitisation element (encoding of audio and scanning of newspapers on microfilm). Although the British Library has more than adequate skills and expertise in-house for these activities they decided to outsource the digitisation because of the large volumes, expensive equipment and tight deadline in which the projects have to be completed.

Outsourcing has its own challenges, particularly at the procurement stage, but in each of these cases The British Library gained significant value for money and high productivity via the outsourcing route. In the case of the newspaper project, this is through offshore scanning, OCR and some keying of content. For the audio project, the ability to convert large volumes of material with an external expert vendor has enabled them to focus internally on their core skills and the difficult issues of Intellectual Property Rights, selection of content and descriptive metadata.

[5] Archival Sound Recordings at the British Library, <http://www.bl.uk/collections/sound-archive/archsoundrec.html>.

[6] 19th-Century British Newspapers at the British Library, <http://www.bl.uk/collections/britishnewspapers1800to1900.html>.

4 Automation

There are a number of ways in which automation can significantly reduce the costs of digitisation. These include:

Mechanistic automation: replacing or reducing the human handling of original materials.

Software based automation: speeding processes, replacing human intervention, or enabling end user interaction that means less effort at creation.

It should always be noted that automation is not a panacea for reducing costs. Just like the Mickey Mouse Sorcerer's Apprentice scene in Disney's Fantasia, automation can sometimes overwhelm rather than improve the workflow. On these occasions, all that is achieved is to move the bottleneck; slowing production to some other part of the workflow, like quality assurance or metadata creation, by delivering more items than can be coped with at current staffing levels.

4.1 Mechanistic automation

When we focus narrowly on the pure costs of image scanning as opposed to the whole workflow there are two cost elements:

a) The cost of handling or otherwise moving the original material through the scanning process.

b) The cost of writing an output image file to the required resolution, bit depth and quality.

For both of these elements the greater the time involved the higher the cost. Technology costs have been reduced for everything but very large files (>300 Mb). This has been achieved through improvements in computer hardware - the use of removable storage, Storage Area Networks and optical fibre networks - such that it becomes a negligible cost as long as it is spread across a large volume project.

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

- A4 laser printed sheets can be passed automatically through a scanner using sheet feeders.
- Bound volumes need every page individually turned on a bookscanner. There are now robotic scanners capable of this for robust originals.
- Photographic prints cannot easily be fed automatically through a scanner because they will be damaged and jam the mechanism.
- 35mm mounted photographic slides can be loaded into a carousel for automated batch scanning.
- Microfilm rolls can be scanned automatically, but microfiche and jacketed film tend to still need human supervision.
- Glass plate photographs can take up to 3-5 minutes each just to get the plate out of its enclosure and onto the scan mechanism, take a scan and then remove the plate back to its enclosure.

Because increased handling means more human intervention then costs will rise. Automation of the transit of materials through a scanner will reduce unit costs.

It is important to note that a low unit price achieved through automation assumes a bureau/production type facility with very high cost machinery to enable large volume imaging. For lower volume projects, unless it considers outsourcing, such high-end scanning equipment might be outside the budget.

The drum fed scanners used commercially, with scan rates of many pages per second and autofeed for several hundred sheets, may easily cost over €35,000. Microfilm scanners often cost more than €50,000 for full automation and greyscale capability.

So, whilst an in-house operation could afford a robust scanner with a 50 sheet feed mechanism for less than a €1,000, the levels of throughput are not comparable and thus costs per image will differ due to scan times being more like 10-30 seconds per scan and the sheet feed needing more frequent reloading.

Robotic scanning

Robotic scanning is a recent innovation that has significantly reduced the imaging cost for bound volumes that are fairly robust. Whilst not suitable for medieval manuscripts, robust bound volumes from the 19th century onwards are capable of being scanned with no human intervention other than the initial placing of the bound volume in the mechanism.

The leading providers in 2006 of robotic book scanners include[7]:

- 4DigitalBooks[8] Digitizing Line claimed capable of turning and scanning 1,500 to 3,000 pages per hour**.**
- Kirtas[9] automatic book scanning range including the APT 2400 claiming 2,400 page image per hour.
- Atiz BookDrive[10], a desktop sized auto-page turning scanner claiming up to 500 pages per hour.

There are also semi automatic bound volume scanners such as provided by Zeutschel[11] and I2S[12] which are human operated but with many features such as auto scan on page turn that they make for high productivity.

All of these are fairly high cost with prices often above €50,000 not at all unusual, but they provide huge potential productivity benefits and savings.

Stanford University has been a leading institution in their application and provide an excellent introduction to robotic scanning of bound volumes. Their workflow explanation is a model for other projects to follow and they provide the following commentary: «actual rates are determined by a book's dimensions, the resolution at which it is scanned, the time taken for occasional operator interventions, and the loading and unloading process… output rates range between 500 and 600 pages scanned per hour. Including metadata entry and manual review of image quality, the average 300 page book can be scanned and converted to a searchable PDF with approximately 40 minutes of operator attention»[13].

4.2. Software based automation

The objective of software based automation is to speed up processes, replacing human intervention, or enabling end user interaction that reduces the effort at content creation.

---

[7] This is not a comprehensive list and new suppliers may join the market at any time.
[8] <http://www.4digitalbooks.com/digitizing-line.htm>.
[9] <http://www.kirtas-tech.com/products.asp>.
[10] <http://www.atiz.com/bookdrive.php>.
[11] <http://www.zeutschel.com/>.
[12] <http://www.i2s-bookscanner.com/>.
[13] Stuart K. Snydman and Catherine A. Aster, *Robotic Book Scanning at the Stanford University Libraries and Academic Information Resources: Report on the Status of Digitization Facilities and Services for Bound Library Materials*, 2003, <http://library.stanford.edu/depts/diroff/DLStatement.html>.

The best understood software based automations are those for batch image manipulation (cropping, deskew, surrogate creation etc.) and character recognition of textual content.

Batch processing has large cost benefits for enabling operators to supervise but not partake in every transaction. The tools to achieve these sorts of batch processes on images can vary from the relatively low cost, such as Adobe Photoshop[14], to the relatively expensive, such as I2S Book Restorer[15].

Automation has yet to deliver large cost reductions in the following areas:

- Picture based metadata extraction (whether still or temporal media like film and video);
- OCR of certain text sources – rekeying can sometimes be more economic (see table below); and
- Intellectual metadata capture – descriptive and context based metadata still generally requires high levels of human expertise and time.

Text capture: Optical Character Recognition and Rekeying

Text capture is a process rather than a single technology. It is the means by which textual content that resides within physical artefacts (such as in books, manuscripts, journals, reports, correspondence etc) may be transferred from that medium into a machine readable format.

The main methods for text capture are (in order popularity of use):

- Optical Character Recognition (OCR) – also known as Intelligent Character Recognition (ICR)
- Rekeying
- Handwriting Recognition (HR)
- Voice or speech recognition

Regarding **handwriting recognition**, Entlich states in his review of the technology 'there is as yet no commercial or open source software for automatic transcription of, or the creation of searchable indexes from, handwritten historical documents.'[16] This guide suggests that handwriting recognition may not be automated in a scaleable fashion unless it has been mediated by forms.

[14] <http://www.adobe.com/products/photoshop/>.
[15] <http://www.i2s-bookscanner.com/en/products_software.asp>.
[16] R. Entlich, *FAQ: Handwriting recognition for historical documents*, "RLG Diginews", February 15, 2004, 8(1), <http:// www.rlg.ac.uk/preserv/ diginews/diginews8-1.htm>.l

Rekeying is basically a human process and thus can be only slightly automated to make it more accurate or quicker. Extremely accurate results can be achieved with rekeying, but it can appear relatively expensive because every character carries a conversion cost and thus the direct costs of capture are very apparent. However, rekeying is generally cheaper than OCR only when the same accuracy is expected (e.g. 99.99% accuracy). This is because correcting and proofreading OCR is more costly than rekeying with cheap offshore services.

Voice or speech recognition still requires a level of human training and optimisation before an automated extraction of the speech to a text format output can be achieved. However, it is proving a difficult technology to make scaleable and cost reducing for digitisation of audio content. This will improve but is not quite mature enough yet to be considered a significant cost reducing technology.

OCR is the most obvious tool for cost reduction in digitisation. It can be used to automatically produce text from digital images and due to the sophistication of fuzzy search engines it no longer has to be absolutely accurate to provide the possibility of very high retrieval rates.

Text Capture Decision Matrix

It is sometimes difficult to choose the right method that will offer the greatest cost reduction for text capture. The table in the following page gives an overview for deciding on an appropriate text capture method.

A tick indicates a useful method and a double tick indicates the most preferred method in terms of cost effectiveness and accuracy. The "*Just type it!"* column is to demonstrate that there are times when it is just easier to get on and employ low cost labour rather then invest time and energy in technical or outsourced solutions.

Key

Simple = Very clear, cleanly printed text in a single column. One language only with no scientific notation, small font sizes, unusual characters/words, tables, graphics or illustrations.

Noisy = Same as Simple except the printed text is not clear or clean because of factors such as dirt, tears, foxing, other marks, creases or show through.

Complex = As Simple but includes either multiple columns, multiple languages, scientific notation, small font sizes, unusual characters/words, tables, graphics or illustrations.

Modern = Post 1950's printed text from a book or journal whose content is mainly black and white with some greyscale or colour

Historic = Pre 1900's printed text from a book or journal whose content is mainly black and white with some greyscale.

Scenario	No & type of page mages	Just type it!	OCR no correction	OCR corrected	Rekeying
Full text or indexing	<100	✓✓			
Indexing: modern	Any volume or type		✓✓	✓	
Indexing: historic	Any volume or type		✓✓	✓	
Full text or indexing for handwriting	Any volume or type	✓			✓✓
Full text: modern	Any volume or type			✓✓	✓
Full text: historic	<1000 simple			✓✓	✓
Full text: historic	<1000 noisy			✓	✓✓
Full text: historic	<1000 complex			✓	✓✓
Full text: historic	>10,000 simple			✓✓	✓
Full text: historic	>10,000 noisy		consider just indexing	✓	✓
Full text: historic	>10,000 complex		consider just indexing	✓	✓✓

Table 1: Decision Matrix for Text Capture

4.3 Case Study: Automating metadata capture

The Metadata Engine Project[17] (METAe), led by the Leopold-Franzens-Universität Innsbruck in Austria, demonstrates how metadata capture and OCR can be optimised and automated to deliver large savings in book and journal digitisation.

METAe makes digitisation easier and more efficient since it detects the structural elements of printed material automatically without any training. These include:

- page numbers and their correct order;
- titlepages;
- table of contents pages;

[17] <http://meta-e.aib.uni-linz.ac.at/>.

- prefaces, appendices, indexes;
- chapters and their hierarchical order;
- issues within journals;
- contributions and their authors;
- running titles;
- illustrations, tables, formulas, advertisements;
- caption lines;
- footnotes; and
- automated double page splitting and cropping.

METAe also demonstrated that metadata can be automatically captured during the digitisation process. Metadata can then be exported as an XML file or PDF file or in any other format suitable to a digital library application. The METAe Engine preferably assembles a METS information object (OAIS) and provides a number of correction tools which allow quality control on all levels.
The METAe engine has been developed in close co-operation with several partners of the project. The partner responsible for the technical development, the German software house CCS-GmbH, distributes the tool as a commercial product under the name docWORKS/METAe Edition[18].
Austrian Literature Online (ALO)[19]: The ALO repository provides free access to more than >10,900 digitised books and journals. The repository is used by several partners:

- University Library Innsbruck
- University Library Graz
- Austrian National Library; and many others.

ALO benefits from the METAe Engine by having:

- Full text searching is more exact than before since searching is only performed on the core content of a book and can exclude the pages found at the front or back of a book, column titles, footnotes, etc.
- Structured searching can be done on the full-text, on the level of chapters and in the case of journals on the level of contributions, on footnotes only, or on caption lines.
- Navigation through the book is easier.

[18] <http://www.ccs-gmbh.de/>.
[19] <http://www.literature.at/webinterface/library>.

- Illustrations, formulas, tables and advertisements become searchable elements of their own right and can be accessed separately from the rest of the content.

Basically, for the same or lower cost huge additional value has been added to the resource through automation of various processes.

There are other suppliers who provide automated metadata extraction or capture. Foremost for the newspaper market is Olive Software[20]. A useful overview of the advantages of this form of XML capture and automated mark-up for newspaper content is provided in an RLG DigiNews article *Digitizing Historic Newspapers: Progress and Prospects*[21].

[20] <http://www.olivesoftware.com/>.
[21] M. Deegan, E. Steinvel, E. King., *Digitizing Historic Newspapers: Progress and Prospects*, "RLG DigiNews" August 15, 2002, Volume 6, Number 4, <http://www.rlg.org/preserv/diginews/diginews6-4.html#feature2>.

5 Selection and Preparation for Digitisation

The selection of originals from a large corpus and the preparation of those originals are significant cost factors that are often hidden or ignored. But these costs will be incurred by every project and thus cannot be ignored.

By far the greatest of these costs is preparation. The following overhead activities will have to be funded, peopled and resourced from within the project.

- Transportation planning for 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.
- Preservation risk management – an assessment of originals to define appropriate transport, handling and digitisation mechanisms.
- The cost of removing the object from its enclosure, removing staples, or cleaning transparencies, or otherwise preparing the physical items.
- The cost of clearing copyright or other rights to use materials.
- inventory check on all original items returned to ensure everything has been returned.
- preservation check to ensure everything returned has not been damaged or degraded by the activity.

All of these activities listed above have to be done or the project becomes exposed to risks of loss and damage to items without means of redress.

It is possible for selection and preparation could account for 20-30% of the total digitisation unit cost. There are a number of steps to reduce or mitigate this cost.

1) Select whole collections where possible as selecting within a collection is time consuming. Selecting small segments from audio or video can be especially costly.

2) Batch originals by their physical nature for scanning. This will reduce the preparation costs as a cohesive workflow designed for that material type can be designed.

3) Organise the workflow so that low cost labour is used and that preparation is appropriate for the scan mechanism.

4) Inventories may be automatically created from existing catalogues and other indexes.

5) Copyright clearance is relatively expensive in time and effort, but cheaper than: litigation; loss of reputation; or having to remove the digital version. A clear procedure and records will make this easier and cheaper.

6) If each original is given a clear and unique identifier then this speeds inventory, makes scanning and metadata capture easier and faster.

6 Increase Performance and throughput

Making the most of the available resources is axiomatic to reducing costs. Digitisation equipment usually deteriorates with heavy use and will depreciate to a very low resale value within 3-5 years. This gives a clear incentive to see equipment being utilised for as much of the available time as possible to gain the maximum benefit from expenditure. One way in which more value could be gained from the current infrastructure would be to increase the available time during which scanning can take place.

Unit costs may be very broadly assessed in equipment and throughput terms as:

$$\text{Unit cost} = \frac{\text{capital expenditure}}{\text{number of items}} + \frac{\text{time expended per item}}{\text{number of items}}$$

Thus, either the available time the scan operation works is increased (throughput) or buy more expensive equipment that can enable automation (capital) and reduce the cost per item.

6.1 Workflow

Streamlining workflows are important to gain maximum throughput for time and capital spent. Workflow is a logically related set of tasks and actions that aim to achieve an objective. Workflow is the collection of work steps and tasks needed to carry out the objective.

Workflow matters because:

- people costs are usually the highest cost;
- barriers or bottlenecks can cost money and time;
- without considering workflow effective planning is impossible;
- risks can be identified before they happen; and
- if it is designed well, the product will be better.

Methods of analysis include looking for the following factors and using them to improve workflow:

Critical paths and routes: identify the steps and the most important actions to be completed and define the path of absolute requirements – this forms the critical path. The critical path should be where most of the resource is being spent – if a peripheral path is seen to take disproportionate more resource it can be reviewed to see if it may be optimised.

The order of occurrence: often by simply changing the order in which things are done can make cost savings.

Parallel activities: identify those activities which can be carried out in parallel and design the workflow to enable this so as to reduce the liner time spent per item.

Inputs and outputs: what is really required and could these be reduced to save costs?

Completion routing: also known as "YES or NO" routing. This can often be automated because the route in which an original or data file is sent depends upon completion of a previous task. This tends to be a source of great automation. For instance, when a watched folder receives new files from the scanner it automatically processes then and saves the output to another folder.

Decision routing: this type or routing assumes some level of judgement and so human interpretation is required.

Calculated value routing: this decision is based on mathematical inputs. For instance, by interrogating the pixel dimensions of a file, the automated data handling protocol could be set to sort thumbnails from master images and place them in separate folders.

6.2 Case Study: National Archives of Sweden

At the beginning of 2003 SVAR Fränsta became a new division in the National Archives of Sweden[22]. The name was changed to Media Konverterings Centrum[23] (MKC). During 2003 and 2004 capital investment of approximately 2 million Euro was made for equipment, software and buildings. MKC now employs over 77 people.

The activity at MKC is concentrated to large scale production of digital images from paper originals and microfilm for different customers. The MKC operate 3 flexible shifts that enable 20 hour working per day and will output ~90,000 images per day (~15 million images per year). The departments are arranged by the various types of originals: maps, bound

[22] <http://www.ra.se/>.
[23] <http://www.mkc.ra.se/>.

books, loose leaf, microfilms and the necessary equipment. This is an extremely effective utilisation of resources and maximises the throughput to reduce the unit cost.

MKC scans hardback bound books and loose sheets in formats from <A5 to A0 in B&W, greyscale and colour. MKC can scan 35mm and 16mm microfilm and also print out the digital images on archival microfilm (16 mm).

A special department prepares the materials for scanning (disassembling folders, bar-coding etc.), which enables using a automatic sheet feeder. The preparation saves up to a tenth of the scanning time.

The working environment is designed with great care to deliver an ergonomic working environment. It shows regard for the operators' daily working hours and conditions to ensure they are comfortable and thus work efficiently.

The MKC is in itself an excellent model for cost reduction, being a governmental system that combines technological services with regard to employment in distant areas. The concept presented by the MKC, central governmental professional services, is in itself an impressive step toward cost reduction. The central investment in infrastructure, the work flow, and the working environment, as mentioned, are also major cost reduction efforts.

7 Continuous Improvement and Quality Assurance

Quality Assurance (QA) that is embedded and a natural part of the activity is a major source of cost reduction for digitisation. Done well it provides opportunities for continuous improvement and optimisation of processes. It needs to be systematic, focussed and proactive not passive.

One mistake often made with QA is to assume it is purely about finding and correcting errors. This is erroneous in that it is unlikely that any project will have sufficient funds to check every single form of output to ensure it has no error of any kind – this would be “gold-plate” the project and be inefficient.

The purpose of QA is to catch errors in process, workflow or human errors. Systematic errors can be corrected by changed the system. For example, if the resolution is often incorrect for the types of materials imaged, then instigating profiles that can be loaded for each type and batching the originals would help to reduce this systematic error. Where human error is found this can be addressed with retraining or by moving that person to a more suitable task for their capability.

QA needs to be focussed upon continuous improvement and optimisation and should never have the purpose of simply producing an error list. Errors should be explored to enable the workflow or process to be changed so that they do not happen again. QA’s aim should be to reduce the amount of rework in a project due to error as rework is very costly to a project.

8 Conclusion

The golden rule for cost reduction is to look at every stage that requires human intervention and either remove it, reduce it or make it as efficient as possible. The most frequently used methods for reducing costs in digitisation are:

1) Reduce the cost of labour;
2) Automate to reduce levels of human intervention in digital conversion and metadata creation;
3) Select and prepare originals to enable higher volumes and reduce variation in the workflow;
4) Increase overall performance and throughput to make the most efficient use of capital expenditure; and
5) Continuous improvement and optimisation through rigorous quality assurance.

Other means for cost reduction include:

- Invest in project management. Good management is a good way to achieve the project in an efficient and economic manner.
- Risk management is a means for cost reduction by anticipating problems and having a plan of action in place.
- Well trained staff can produce higher productivity and better quality for less money.
- Cooperation is a means for cost reduction. Some ideas:
    - Joint projects across a few institutions, dividing the tasks and the costs of the project, thus aggregating benefits and reducing the costs for each partner.
    - Barter scanning or other processes in an exchange with institutions having complementary equipment. This reduces the capital outlay for all partners.
    - Use technology and automation whenever possible to reduce human time and costs.
    - Stick to your core needs and critical requirement and thus avoid project drift or the temptation to "gold plate" projects.

About the author

Simon Tanner has a Library and Information Science degree and background. He is the founding Director of King's Digital Consultancy Services (KDCS) at King's College London. KDCS provides research and consulting services specialising in the information and digital domain for the cultural, heritage and information sectors. He has advised and managed over 450 digitisation projects across Europe and America with libraries, museums, archives and corporate institutions.

Before joining King's, he was Senior Consultant at HEDS - the Higher Education Digitisation Service - and had a key role in its successful development as a JISC Service. He has also previously held IT, management and library roles for Loughborough University (Library Systems Manager), Rolls-Royce and Associates (Head of Library Services) and IBM (UK) Laboratories (Information Officer).

Simon is an independent member of the UK Legal Deposit Advisory Panel and Chair of its Web Archiving subcommittee. He has been a consultant to UNESCO, the National Library of Scotland, Kew Gardens, the Royal Academy of Arts, Imperial War Museum, Oxford University, the National Library of Ireland, Birmingham Public Libraries, the House of Commons and advised Denmark’s national digitisation programme amongst many others. He has also carried out research projects for the Andrew W. Mellon Foundation on charging models for digital cultural heritage in Europe and the USA.

Simon authored the book, Digital Futures: Strategies for the Information Age, with Dr Marilyn Deegan.

Printed by
La Tipografia di Umberto Frisardi s.a.s.
Rome, Italy
in the month of October 2006