Outsourcing

IMPACT Best Practice Guide

IMPACT project
Gerd Zechmeister, Austrian National Library
Mark-Oliver Fischer, Munich Digitization
Center at the Bayerische Staatsbibliothek
Niall Anderson, British Library
Neil Fitzgerald, British Library

Table of Contents

A Best Practice Guide to Outsourcing

Introduction

Museums, Libraries and Archives have been digitising their collections for more than twenty years, but industrial digitisation – the conversion of millions of information resources into digital form – is still a relatively young field. For that reason, many institutions lack the internal knowledge or physical resources to manage large-scale digitisation projects without outside assistance.

Depending on an institution's internal capacity and policies, a digitisation project may require third-party input at any stage of its workflow: from initial image capture and storage, to metadata creation and quality assurance of project outputs. In such cases, outsourcing services or tasks to a third-party can be greatly beneficial in terms of scoping, managing and costing a digitisation project from beginning to end.

While outsourcing may deliver a more streamlined and cost-effective workflow, such collaboration can also bring potential complications: for instance, the necessity of defining the legal ownership of the resulting work, the possible use of proprietary software in creating the image files and metadata (which may limit the reuse potential of project outputs), or restrictions on access for commercial reasons.

Main benefits

Large-scale digitisation projects may require input from external suppliers at any stage. For instance, a digitising institution may lack the physical capacity to undertake the necessary work in-house, or there may be a lack of knowledge within the institution about how to digitise very complex material (e.g. newspapers) cost effectively. In such cases, outsourcing tasks to a third-party can be greatly beneficial in terms of scoping, managing and costing a digitisation project from beginning to end.

Main challenges

While outsourcing may deliver a more streamlined and cost-effective workflow, such collaboration can also bring potential complications: for instance, the necessity of defining the legal ownership of the resulting work, the possible use of proprietary software in creating the image files and metadata (which may limit the reuse potential of the work), restrictions on access for commercial reasons, etc.

Outsourcing a part of the workflow also means that the digitising institution is less likely to retain working knowledge of that part of the workflow for its next project.

Background and developments to date

Outsourcing parts of a digitisation workflow has become increasingly common practice since new technologies (in particular, automated scanners that can capture hundreds of images a minute) made industrial digitisation of text resources a viable practice.

Although investments in technology and infrastructure should be considered as an option for large-scale digitisation projects, service providers may come up with more attractive pricing models which can result in lower costs. A critical decision factor on the other hand is capacity building in-house, with outsourcing the digitising institution is less likely to retain working knowledge of that outsourced workflow for its next project. Long-term considerations to build capacity (and therefore investing into resources) to realise entire projects or parts of it in-house should be considered as an option in accordance with institutional digitisation strategies.

In the last few years, Google, Microsoft and others have worked with major libraries across the world to provide online access to digital copies of collection material. Alongside initiatives such the European Commission's i2010 Declaration (which envisions digitisation as a core activity of libraries and archives, through the creation of a "critical mass" of online digital content), these huge undertakings have effectively raised the bar of what is possible – and expected – from digitisation projects worldwide.

Projects such as Google Books differ from the common run of digitisation projects in that practically all stages of the work are managed outside the structures of the library or archive whose content is being digitised. In this respect, they represent a particularly radical form of outsourcing.[1] However, the vast majority of digitisation projects still take place under the direct management of the parent institution, with that institution making decisions about which parts of the workflow to outsource and which to manage in-house. This document addresses itself primarily to those institutions, while acknowledging that the kind of total solution offered by Google and its associates has implications for all digitising institutions. It gives practical advice on how to assess whether an institution needs to subcontract some or all of its digitisation work to a third-party, how to engage with and assess service providers before work begins, how to establish and manage the relationship with the company, and how to retain within the digitising institution practical knowledge about digitisation gained in the course of working with service providers.

---

[1] For an in-depth look at the workflow devised between Google and a participating library, please see: *Mass Digitisation Workflow Management;* B.I.T. Online; 2008; http://www.imageware.de/static/common/files/en/156/BIT_3_BSB_Beitrag_engl.pdf Retrieved 13.03.2011

Best practice in managing outsourcing

A digitising institution should have three main considerations when considering outsourcing:

• What services or tasks of the workflow shall be outsourced?

• How to contact, assess, and select appropriate candidates for outsourcing projects?

• How to manage the project with and establish good relations to the service provider?

The following chart outlines these considerations by translating them into six main phases and breaking them further down into subtopics. They are outlined in the following pages:
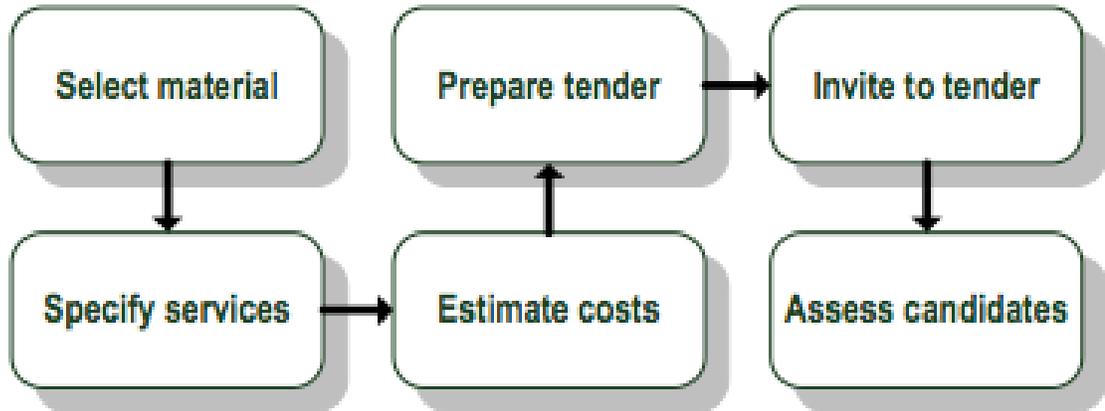
Phase 1: Select the material

Work out a detailed and well-structured overview of the quantity and characteristics of the material to be processed.

• Provide exact numbers (e.g. a batch consists of 500 pages) or reasonable and comprehensive estimations of quantities (e.g. the average number of shots per microfilm jacket is 20, average pages per issue is 30) in a quantity structure. Pilot stage digitisations of representative sample material might be helpful to determine average numbers of pages. Estimations can also base on physical shelf lengths or weights of material. Make sure that your calculations are reasonable and comprehensive;

• For OCR processing be very specific in estimating the amount of characters to be transcribed (e.g. average of 8000 characters per page) and what quality you require. Word or character accuracy rates (expressed in % of correctly transcribed characters) are the most common metrics to measure the quality of output;

• Provide excerpts from your institutional resource discovery plan to inform the service provider on how the project output will be applied;

• Inform the service provider about pages or volumes in poor condition (e.g. brittle pages) and consider how to capture this type of added value information from the service provider while they scan them;

• Check if the material is allowed to leave the premises, or in the case of uncut pages, are they allowed to be cut and by whom?

• Provide documentation to address your requirements regarding the insurance of material, the offsite storage environment and your security policies;

- Make clear agreements about quality assurance and assesment of the results.

Phase 2: Specify services

Describe the work to be carried out and determine what should be outsourced and what should be done in-house. Mark the interfaces where in-house services or tasks stop and out-house services or tasks start (e.g. between "transporting books to service provider" and "preparing books for scanning"). A clear and comprehensive description and documentation of work can greatly cut down issues and misunderstandings between parties, as well as clarifying the responsibilities of each party in case of a breakdown in the workflow or a change of scope in the project.

Typical services for digitisation and OCR projects are document capture/digital imaging (including specifications of masters and derivatives), OCR processing (OCR standards and output format such as full text, ALTO, PDF, etc.) and metadata identification and extraction (how metadata is identified, added or extracted from a database, and in what form).

Consider what expertise and skills are available in your institution and which should be acquired. Consider also the necessary quality standards of the project: does the institution have experience of quality assurance procedures, or a ready-made quality assurance system? In the absence of such a system, quality assurance usually requires human oversight and can be a high-cost factor to the institution.

The quality of output shall be assessed and/or measured by predefined metrics: where OCR is to be applied, correction of OCR should be specified in relation to document parts: full text, title, article headings etc. Some digitisation projects may only need an article heading to be recognised with an accuracy rate of nearly 100% correctly at all times while body texts don't need to be quality-controlled manually when a pilot delivers satisfying results.

A small but representative pilot of the work to be outsourced is strongly recommended before full production begins, to identify and mitigate possible issues. A predefined set of test material to be processed by the candidates shows the expertise required and the expected output. If a pilot is not intended, prepare some representative sample material to be assessed by the service provider.

Phase 3: Estimate costs

A table of anticipated costs should be compiled to give a clear overview of the work and the spread of costs throughout the project. It is also a means by which various third-party service providers can be assessed against each other. Keep in mind that the complexity of processing strongly influences prices and models. Definitions and descriptions of used terms (e.g. how do you calculate the accuracy rate?) shall be attached.

Example table for digitisation and OCR services of books:

One-off positions	Quantity	Units
Process setup	1	service
Running costs positions	**Quantity**	**Units**
Image capture: Master image production	1	page
Image optimisation: application of basic image-manipulation techniques (de-skewing, cropping, rotation)	1	page
Derivative Image production (JPEG)	1	page
Metadata file creation (METS)	1	book
OCR processing: creation of raw full text output file (TXT)	1.000	characters

Manual correction of raw full text OCR output file (Minimum accuracy rate: 99% at character or word level): creation of corrected OCR output file (TXT)	1.000	characters
Detection of chapter headings and creation of table of content	1	book
Automatic data validation, file naming and delivery folder assembly	1	book

Find further information on budgeting for a digitisation project here: http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/budgeting-for-a-digitisation-project/ Retrieved 13.03.2011

By the end of this planning process, the following information should have been largely or entirely clarified:

- Project key data (quantities of input documents, description of services and tasks, timeframes for starting and completing work etc.);

- Technical requirements (master and derivative formats, resolutions, colour depth, Metadata and OCR output file formats, file naming conventions etc.);

- (Preliminary) workflow (outlining the processes of preparation of material, exchange of digital files, quality control, automatic data validation etc.) – a graphical workflow can be accompanied by text oulines (e.g. quality control: documents and images are received, inspected and controlled by the quality assurance team. Results matching the determined quality criteria will be approved, non-matching results will be rejected and have to be corrected or reprocessed by the service provider);

- (Preliminary) Quality Assurance Plan (determination of randomly selected n% of results to be controlled, how much time and staff it requires etc.);

- Sample material (e.g. categorised in good/medium/poor quality sources; if necessary with corresponding destined output files) and/or test material with corresponding processing instructions/technical requirements;

- Cost estimation table (for internal use only);

- Resources and capacities available in-house and approved being assigned to/used in this project: key staff, hardware, software etc.;

- List of prioritised vulnerabilities/potential risks and planned mitigation measures and escalation procedures;

- Assigned key staff and their planned roles/responsibilities in the project (e.g. the project manager coordinates the work and documents the progress of the project; the quality assurance team checks the quality and feedbacks to the project manager);

- Transportation regulations and insurance policies of fragile and unique material

Phase 4: Prepare tender

Consult experienced institutions, partners and experts for their knowledge of reliable companies offering the services you require. If an institution is planning to contract outside its main economic area (e.g. the European Union), it should find out if that service provider has a local branch and conduct work primarily through that branch and/or named single point of contact.

If a public tender is mandatory the service buyer must apply to corresponding tender rules in its current form.

Find further information on legislation in the Information system for European public procurement: http://simap.europa.eu/buyer/links-buyer/index_en.htm

Phase 5: Invite to tender

Depending on the complexity of work or the assumed project volume it might be necessary to contact candidates via a request for information (RFI) prior to a request for proposal or quotation (RFP/RFQ) to identify appropriate candidates. Candidates are requested to deliver a detailed project plan which shall at least contain information on production capacities (e.g. maximum number of files/jobs processed in a given time) and a delivery plan including exact dates and quantities.

Beside the project plan most tenders require candidates to submit specific company information. The list of assessment criteria listed in the next section can be used to assemble forms to receive the correct information – though if the amount of work to be subcontracted is small or focuses on a very specific part of the workflow, certain documents and information may not be necessary.

Phase 6: Assess candidates

The assessment process must comply with legal regulations and might be supported by selection criteria arranged in a tender scoring or evaluation matrix. Based on the results a shortlist will determine the best-bidding and/or most appropriate candidates.

Sample matrix:

Criterion	Weight	Score	Weighted score	Comments/evidence
	(0 to 5)	(0 to 5)	(weight x score)	
Tenders are expected to give a clear demonstration of their skills and experience by running a pilot	5	3		Clear understanding of requirements Quality criteria met for less than 80% (threshold) of the material processed
…	…	…	…	…

Explanation of points: 0 Points - Unsatisfactory response, 1 Point - Less than satisfactory response, 2 Points - Satisfactory response, 3 Points - Above satisfactory response, 4 Points - Very good response, 5 Points - Response exceeded expectations

Following criteria can be used as starting point to define selection criteria and can be adapted to specific needs:

Formal criteria:

Eligibility	Are there restriction of companies participating in the tender (e.g. companies with European offices only; work has to be accomplished within European Union)?
Date of submission of tenders	Did the candidate submit the tender documents within the given deadlines?
Form of tenders	Did the candidate include/fill in all necessary documents in the required form?
Number of tenders	Is there a minimum/maximum number of tenders allowed?

Proposal assessment criteria:

Price offers	Specification of running (prices per service, page/document/article/book, processing of thousand characters, per person hour/person day etc.), one-off costs (training, setup, software development and/or parameterisation etc.) and estimated total costs
Commercial and general terms	Prices inclusive/exclusive of VAT, terms of validation of tenders
Project plan	Check if the delivery dates and quantities fit to the requirements set up in the planning phase
Test set results	Evaluate the test sets according to the quality criteria and predefined metrics
Material handling	Transportation and storage of material (especially fragile, unique and loose/unbound material), implementation of a tracking system by annotating items (one book, one batch etc.) with current status ("on the way to provider", "in preparation for scanning" etc.)

Company assessment criteria:

Economic position	Annual turnover made with services required in the tender, financial state of the company etc.
Affiliations and registrations	Chamber of commerce, proof of registration in the national professional register etc.
Alliances and partnerships	Cooperation with other providers to deliver services or tasks or a software company develops an online quality management system etc.
Certificates	Awarded/evaluated for applied quality management system, environmental and social certificates etc. – check the date of evaluation and the integrity of the awarding authority.
References of work	List of previous similar projects with descriptions of services, date of accomplishment, earned turnover, number of employees involved etc.
Application of standards	Project management, quality assurance etc.
Expertise	Experience and certificates of key staff such as production or project managers (e.g. education, skills, qualifications)
Security	Security and data protection policies (including business disaster recovery and contingency plans), policies on handling of fragile and unique material

Find standard forms for public procurement in the Information system for European public procurement: http://simap.europa.eu/buyer/forms-standard/index_en.htm

Following documents and information should be provided by the contracting institutions to start negotiations:

• Draft agreement/legal contract (including articles about definitions, terms and conditions, warranty, prices and payment, acceptance, services, confidentiality, intellectual property rights, documentation, working hours, liability, transfer of rights and obligations, insurance, termination, force majeure, disputes, applicable law and choice of domicile, service level agreements, non-disclosure agreement, escalation procedures and penalties);

- Change-request template: requested changes of requirements/focus/scope of project must be agreed on mutually by contracting institution and service provider;

- Delivery report template (incl. date of delivery, quantities of delivered files, report on errors etc.).

Establishing and maintaining good relations in Outsourcing projects

The scenario of working together in a business relation strongly depends on the complexity level of the required services and the special know-how and experience service providers offer.

It is in the nature of outsourcing that one party is effectively employing another. Nevertheless the process should ideally be seen as a cooperative partnership. Good relations between the parties should be considered as common goal for successfully reaching common goals and – if intended – establish long-term cooperation (e.g. in form of service provision contracts).

The service provider should not just supply a defined and limited service, but also provide expertise and recommendations to improve the quality of results. The digitising institution as contractor on the other hand shall deliver clear and consistent information and instructions on the services required.

The success of projects is not only crucial for the adoption and application of the output but also for the setting of upcoming projects. A common understanding between key persons of both parties and well-rehearsed processes may save time and money when impossibly preparing future projects.

Being a good customer: What do service providers expect from their customers?

In 2007, OCLC Programs and Research engaged Intelligent Television to study the partnership agreements between cultural institutions and for-profit companies for the industrial digitisation of books and other media. The report made the following seven recommendations about how non-commercial entities such as libraries should forge an understanding with commercial suppliers (US English is preserved in this fragment):

> 1. Recognize that the business models of companies in digitization deals today are often fresh iterations of longstanding business models espoused by companies who have long been active in the field, such as ProQuest and Readex NewsBank. Those arrangements often resulted in copies of original library materials being provided to their partner libraries (and revenue generated from the sale of those products was sometimes shared with the library that contributed the original materials).
>
> 2. Recognize, equally, that some of the business models are entirely new, reflecting the fact that these commercial enterprises are now producers or co-producers of screen-based media in the digital age and much of that media is now advertiser-supported, provided by paid subscription, or on-demand.
>
> 3. Recognize that commercial companies seek returns on their investments in Industrial Digitisation based on economic calculations including:
>
> a. effects on near-term revenue;
>
> b. effects on closing future deals that in turn may bring in additional future revenue;
>
> c. effects on corporate profit;

d. effects on closing future deals that may bring additional profit; and

e. effects on company valuation

and that these calculations may tend to be more short-term than those of cultural and educational institutions.

4. Consider that the investment banks and management and strategy firms that appraise and influence the valuation of business partners are paying increasing attention to new ways of measuring value in the online world. Rendering content searchable, or what some have called "computably competitive," may be as good an investment of resources, if not better, than simply making more content available.

5. Recognize that private companies often need a competitive edge in return for their investment. They desire to accumulate content that their competitors do not have; they want some measure of exclusivity to allow their investments to return value; and they want to protect their plans and approaches to technology and media.

6. Recognize that businesses see cultural and educational institutions playing a critical role in the $12 billion annual business of online search—not only as a customer and as a gateway to these online services, but also as a potential competitor.

7. Finally, recognize that while commercial companies are bringing to negotiations and business deals their billion-dollar valuations and hundreds of millions of dollars in capital to support digitisation, they may not recognize that cultural institutions themselves represent cumulatively billions of dollars of investment, based on the value of their assets and decades (if not centuries) of collecting, curating, and preserving physical copies of these works.

In strictly practical terms, what service providers need above all is clarity from the digitising institution. The institution should focus on three areas in the workflow and give specific and detailed instructions:

- Input: all the material the institution will be submitting for processing (e.g. books, digital files), structured by number of complete issues (e.g. entire books or newspapers) or batches thereof (e.g. divide unbound pages belonging together into packages of 500 single pages).

- Process: state clearly any special necessities or restrictions on the workflow suggested by the service provider. For instance, if a book has uncut pages, is the service provider allowed to cut them?

- Output: the results which the service provider has to deliver. Define the technical requirements such as file format, data structure, character encoding, file naming conventions etc. Some assignments require more iteration and might produce intermediate results: Scanned Images # OCR processing # raw OCR output in plain text (output 1) # quality control and tagging/conversion # XML file (output 2). Select representative examples and provide input and output files the way you expect them (e.g. plain text files encoded in UTF-8) and carry out tests to have the output and quality clearly defined.

Conclusions and future developments

The future of outsourcing as a viable practice in industrial digitisation lies not in the rigid application of a particular workflow or system as much as it does in fostering greater flexibility and understanding between commercial and non-commercial partners. Evidence of the Internet Archive/Google scanning projects is that they do apply a one size fits all approach which can limit choice, while the possible benefits and risk of outsourcing is that you can customize the workflow. The ability to modify can have a significant impact on costs reducing outsourcing benefits.

The OCLC/Intelligent Television report cited above made recommendations as to what libraries and archives should aim for when negotiating contracts with commercial parties. These recommendations were

made in the light of a number of large digitisation projects – delivered partly through outsourcing, on time and on budget – which were delivered to contract, but whose long-term results for the library partner were less than successful. The library or archive should seek to maximise its return on a project by seeking:

- **Limited confidentiality** – Libraries should be sensitive to private partner needs to protect business and technology secrets, but insist on their own right to discuss aspects pertaining to their broader community. These deals involve some of the most complex decisions libraries will face, they can be improved through consultations with others, and libraries do not want to be in the position of having to refuse advice to peers who seek their guidance.

- **More complete deliverables** – Librarians must have input into the specifications of quality and formats and be clear about exactly what they will receive. They must ensure that they will own those deliverables or they remain in the public domain.

- **More open access** – Librarians should preserve their right to provide unrestricted access to users. In particular, they should avoid contract terms that make it difficult or impossible to offer scholars the kinds of functionality, including automated or bulk access to collections that can support innovative research and will allow the development of new applications.

- **Less restricted distribution** – Librarians should preserve the right to combine parts or all of their digitised content with collections at other institutions or non-profit organizations.

- **Responsible treatment of usage data** – Librarians should ensure that users' privacy is protected, even while drawing on usage data to enhance services to users.

If the above terms cannot be secured, then the consequences of compromises should be fully understood. This point, however, should be a requirement:

- **Limited duration and survivability** – Restrictions on ownership, access, and distribution should not survive termination of the agreement.

When entering into an outsourcing agreement, it is important for both partners to remember that digitisation is a form of preservation as well as a means of remote access to material. If rigid or proprietary standards are applied to any of the technical processes in a workflow, there is a risk that the use of the digital resources will be limited to the lifespan of the project that created them. This is something that museums, libraries and archives, in particular, should seek to avoid.

References

Kaufman, P.B., Ubois, J. (2007). Good Terms – Improving Commercial-Noncommercial Partnerships for Industrial digitisation. *D-Lib Magazine, 13(11/12).* http://www.dlib.org/dlib/november07/kaufman/11kaufman.html Retrieved 12.03.2011

JISC Digital Media, Cross-Media, Budgeting for a Digitisation Project. http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/budgeting-for-a-digitisation-project/ Retrieved 12.03.2011