



IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Challenges and Opportunities in Mass Digitisation

How Technology Can Meet Libraries' Needs

Apostolos Antonacopoulos and Stefan Pletschacher
PRImA, University of Salford, UK

7 April 2009 presentation The Hague



Goals of discussion

1. Facilitate a common understanding of challenges faced by libraries and by technology researchers
2. Elicit objectives from the libraries, in view of what can be reasonably expected from current/projected technology
3. Map those objectives to measurable technological goals



Outline

- Prior to the conference we sent out a questionnaire to libraries and service providers to gather information as a starting point to this discussion
- The following slides present a draft overview of the responses received (a report containing this information and points raised during the discussion will be made available in due course)



Material

- List the types of material in your library that could be part of a mass digitisation programme (now or in the next five to ten years). Rank them in terms of volume (e.g. Newspapers – 10 million pages, books 19th century – 5 million pages etc).
- How do you prioritise what will be digitised in the near future? Please provide examples of material (type, age, language etc.)?
- Provide some representative examples of your material which is likely to be digitised (include links to images or bring a CD/USB stick to the conference).



Material - Types

- Books
- Newspapers
- Journals, magazines
- Parliamentary and legal papers
- Patents, technical reports
- Manuscripts
- Maps
- Sheet music
- From Originals as well as Microfilm
- Up to several millions
- Mostly before early 20th century – out of copyright, less modern books



Material - Priorities

- Cultural heritage (national literature etc.)
- Unique and valuable items
- Deteriorating material and items in danger (bad physical state - paper, faint and vanishing text etc.)
- Customer requirements, most requested (e.g. newspapers etc.)
- Value for money! Material with the best ratio of effort to result (highest expected outcome for as less money as possible)
- Out of copyright!
- Funding possible
- Easy to difficult
- Everything!



Material - Examples

- Please send us material if possible
- Links to specific images (rather than collections) ... we know that it is difficult...



Use

- How is your already digitised material used by readers (Full text search, image display, PDF files etc.)?
- What are planned use scenarios of digitised material (Web, full text search, re-purposing, editing, output on mobile platforms, re-print etc.)?
- In an ideal world, if there were no technical or financial obstacles, what do you think would be the best way of making available the digitised/full-text material for use by the readers (e.g. full-text access through mobile devices)?
- How important is it to improve the visual appearance and readability (for humans) of the original image itself for better viewing? For instance, would it be useful for an enhanced (with defects, like page curl, corrected) version of the original document image to be presented to the reader?



Use – Current Usage

- Usually available through a web-based portal to the database
- From full-text search to table of content search to metadata-only search
- Viewer for images
- Some libraries also provide download facilities for files
- PDF files (image only, searchable – image and hidden text, corrected or “dirty”, also pre-print versions from publishers in some cases)
- DjVu
- ALTO, TEI, different XML formats as well as plain text
- JPEG, JPEG2000, PNG (sometimes different resolutions available)
- TIFF (often not accessible for users)
- METS for metadata



Use – Planned Use Scenarios

- Web access
- Full text search
- Full text access
- Re-purposing, re-use, annotation, editing
- Re-print, print on demand
- Off-line media
- Word highlighting, contextual search
- Output and usage on mobile devices
- Download facilities
- PDF as well as export formats for exchange
- Control of access (digital rights management)
- Web 2.0 ???
- Personal virtual book shelves
- Semantic analysis of text and extended search functionalities (e.g. find corresponding material – sophisticated information retrieval)



Use – The Ideal World

- Access to all material
- From everywhere
- Anytime and fast
- On any platform, device (mobile devices, e-books etc.)
- Online as well as download
- For free :-)
- 100% correct text, handling of spelling variants
- Metadata
- Rich information network with links to other information sources like biographies, encyclopaedias etc.
- Handwriting recognition
- Mathematical and chemical expressions
- Clear Copyright regulations



Use – Importance of Visual Appearance

- No
- Rarely
- Nice to have
- Very important

- Majority of the users are interested in information and not in the visual appearance
- If it improves OCR
- As addition to un-enhanced images
- If it could be done completely automatically
- Could enable new use scenarios (e.g. re-print)
- Depending on “business model”



Issues

- What are the steps in your digitisation pipeline?
- Which processing steps cause the highest effort/cost (e.g. scanning, image enhancement, OCR, post-correction etc.)?
- Which processing steps, in your experience, are currently not satisfactory in terms of supporting tools and/or results produced (e.g. image enhancement, segmentation of newspaper articles, OCR of archaic fonts etc.)? Suggest possible improvements.
- In your experience, what have been the biggest problems in implementing/completing digitisation programmes?



Issues – Digitisation Pipeline

- Selection
- Copyright clearance
- Collection of the physical object
- Preparation (including restoration if required)
- Metadata of physical objects
- Transport to supplier →
- Ascertainment of scanning parameters
- Scanning
- Quality assurance
- Image enhancement
- Quality assurance
- Segmentation
- OCR
- OCR post-correction
- → Delivery of digital files
- Quality assurance
- Indexing (metadata for digital objects)
- Ingest, storage
- Linking to other electronic catalogues or databases
- Publication on website
- Production of alternative output formats (re-formatting etc.)
- Long term preservation, digital archiving



Issues – Highest Effort / Costs

- Strongly depending on the implemented workflow
- Preparation of scanning
- Scanning
- Image enhancement
- OCR
- Post-correction
- Quality assurance in general
- Time: capture and input of metadata
- Adhoc treatment of documents
- Handling of multilingual material
- Manual typing...



Issues – Not Sufficiently Supported

- OCR of archaic/non-standard fonts (Gothic etc.)
- Omnifont recognition
- Multilingual texts, non-western material, special dictionaries
- Segmentation
- Non-standard layouts, tables, charts, maps, formulas
- Image enhancement (show-through, poorly printed documents, straightening of text lines)
- Automation in general
- Automatic extraction of metadata and document structure
- Quality Assurance - tools and benchmarking needed.
- Tools for managing and tracking projects, workflow support
- Clearance of Copyrights



Issues – Implementation of Programmes

- Coordination, time management
- Heterogenous projects
- Money and effort
- Lack of continuity in funding of digitisation projects
- Persuade government authorities to constantly support digitisation projects
- Manpower, staff training
- Definition and implementation of workflows
- Distributed infrastructures
- Metadata
- Getting what you want from suppliers
- Copyright issues
- Recognition rates
- Fonts and languages



Evaluation

- How do/would you measure the success of a digitisation programme? Please provide any measures currently used and how those measures are derived.



Evaluation – Measuring Success

- Users!
- Acceptance by users
- Number of unique visitors per day
- Feedback from users
- Coverage – how many different groups of users benefit from the end result
- Critical mass
- Number of pages produced per year
- Delivery according to work plan
- Completeness
- Finishing projects on time, with the given budget
- Quality of images (scans)
- Legibility of scans
- Recognition rate
- Percentage of correctly found documents
- Level of automation
- Comparison to competitors
- ... no formal analysis
- ... still working on formal measures



Contact

Apostolos Antonacopoulos, Stefan Pletschacher



www.primaresearch.org