Mass Digitisation QA

IMPACT Case Studies

IMPACT project
Geneviève Cron, BNF - Bibliotheque nationale de France
Neil Fitzgerald, British Library
Niall Anderson, British Library

Table of Contents

Case Studies of Mass Digitisation QA

The two examples below provide an illustration on how to integrate a quality procedure into the workflow of a digitisation programme:

Case study: Quality control system for mass digitisation at the Bibliothèque National de France

Automated checking

Once the digital document (images files, OCR files, table of content file, metadata file) is delivered by the contractor, the IT department uses a routine for an automatic check of the document and on each of its files. This process checks the header of each TIFF file, the structure and content of XML files (ALTO, metadata, Table of Content) and the identifier of each part. IDs are checked to ensure that the digital document will be linked to the bibliographic record in the catalogue and that the package submitted to the archiving system allows the building of a coherent archival package.

The programme returns a log file containing the list of errors, and some warnings (potential errors) which are checked manually. The results are stored in a database to monitor the level of correction and the delivery of the improved document. The history of corrections done on a document is indicated in the metadata.

Publishing and auditing

When the document is validated by the automatic control, it is directly published on Gallica (however just a part of the payment is done to the contractor). An audit system is in place to verify each process of

the production line of the provider and on the conduction of the project. The audits take place every two months and are based on visual inspection of samples. The BNF has tried to limit the risks by taking time to validate the quality produced during a long test period. Audits can anticipate the risk of errors on the upcoming documents by intervening directly in production lines. In addition, the BNF may request the repair of documents already online for a period of one year after the automatic admission.

Case study: Microsoft Digitisation Project (MDP) at the British Library

For MDP, once an item was scanned automated checks were applied immediately by the contractor to identify possible errors and resolve immediately. These automated checks include verification of resolution, bit depth, file format, file dimensions (width and height, in relation to the average dimensions of each book), and file size (in kilobytes, in relation to the average file size of each image).

At the capture stage

Any non-conforming suspicious value resulted in a raised flag and called an operator to review the particular set of images and to decide if a problem existed. If there was a problem and it could be fixed, the operator would take the appropriate action. If it could not be fixed, the operator could reject the book, logging the deselection in the book-tracking database for later reports. If there was no problem, the operator sent the digital book to further processing. All of the actions taken were logged into the book-tracking database for later reports and statistics.

Remote QA stage

Once the file integrity had been established, the files were subject to a number of automated checks. These included establishing that the barcode under which an item had been retrieved matched the details of that item's BL catalogue record retrieved through z39.50; checking for blank, unnumbered or out-of-sequence pages in a file (which may suggest errors in the original scanning or pagination errors); an OCR "confidence" algorithm; a crosscheck of OCR results with recognised words from a dictionary; and for the investigation of text blocks for which there was no OCR result. A negative result in any of these fields would lead to the item being flagged up for an operator to investigate, rescanning or deselecting if necessary.

After all checks and corrections are done, METS/ALTO xml metadata files are created and saved in the output directory. Image files representing each page are generated as well. In creating this output, several parameters are stored and validated: image format, image resolution, file size, bit depth, and a list of all files generated from a single volume.

Finished files are delivered by the contractor on at least a weekly basis. Each delivery was accompanied by a batch manifest[1] stating the precise amount of books and pages and some more statistical data per book. Quality Control using the modified ISO 2895-1 standard was then carried out by the BL, logging any major or minor errors in the batch manifest for items to be reworked. The expectation was that more batches would pass than be rejected and since full production has been established it is now very rare for any items to have this status. The project's metadata schemas/file integrity was validated by the JHOVE validation tool[2].

Batch manifests [http://www.bl.uk/schemas/deliveryManifest-v1-0.xsd] were xml documents with the following structure and functions:

[1] Data Exchange inside the Microsoft Digitisation Project; British Library and Content Conversion Specialists GmBH; 2006: http://www.bl.uk/schemas/deliveryManifest-v1-1.xsd Retrieved 13.03.2011

[2] JSTOR/Harvard Object Validation Environment; 2009; JSTOR and Harvard College: http://hul.harvard.edu/jhove/ Retrieved 23.03.2011

ILSID	The unique ILS ID for the book "003448648"
barcodeID	The barcode ID from the ticket "A0010219365"
in	Date the book has been delivered to the scan studio
scan	Date of scanning
out	Date of delivery to MS/BL
pages	Number of Pages
sourcetype	Type of book (Multivolume/ or Single Volume)
code	Code according to specification
ocr	OCR Confidence value (average for the book)
foldouts	number of foldouts included in the particular book
status	Status (PREPARED, PROCESSED or NEED ACTION)
	PREPARED – provided by CCS, to be verified by MS/BL
	PROCESSED – verified/accepted by MS/BL
	NEED ACTION – not accepted by MS/BL after verification
body	Whenever there is a reject and rework request identified in the quality assurance process of Microsoft or British Library, the specific comment is written in the body of the XML tag.

Books were assigned NEED ACTION status by the BL/Microsoft if they did not conform to the Acceptable Quality Level established for a scan of a particular type of source material. In the following table, each type of error is assigned a numerical value. A total result above that agreed as Acceptable would lead to the individual item/batch being rejected:

Attribute	**AQL**	**Unit of Measure**
Automated Data Testing	1.0	Various file formats
XML File Inventory	2.5	XML file
Image Skew [SKW]	1.0	Image file
Image Crop [CROP]	1.0	Image file
Image Quality [QUAL]	1.5	Image file
Image Sequence [PAG]	1.0	Image file
Duplicate or missing image [DUPP] or [MIS}	1.0	Image file
Missing or Poor OCR result [OCR]	1.0	ALTO XML file
Missing PDF file or PDF quality issue [PDF]	1.0	Bound PDF file

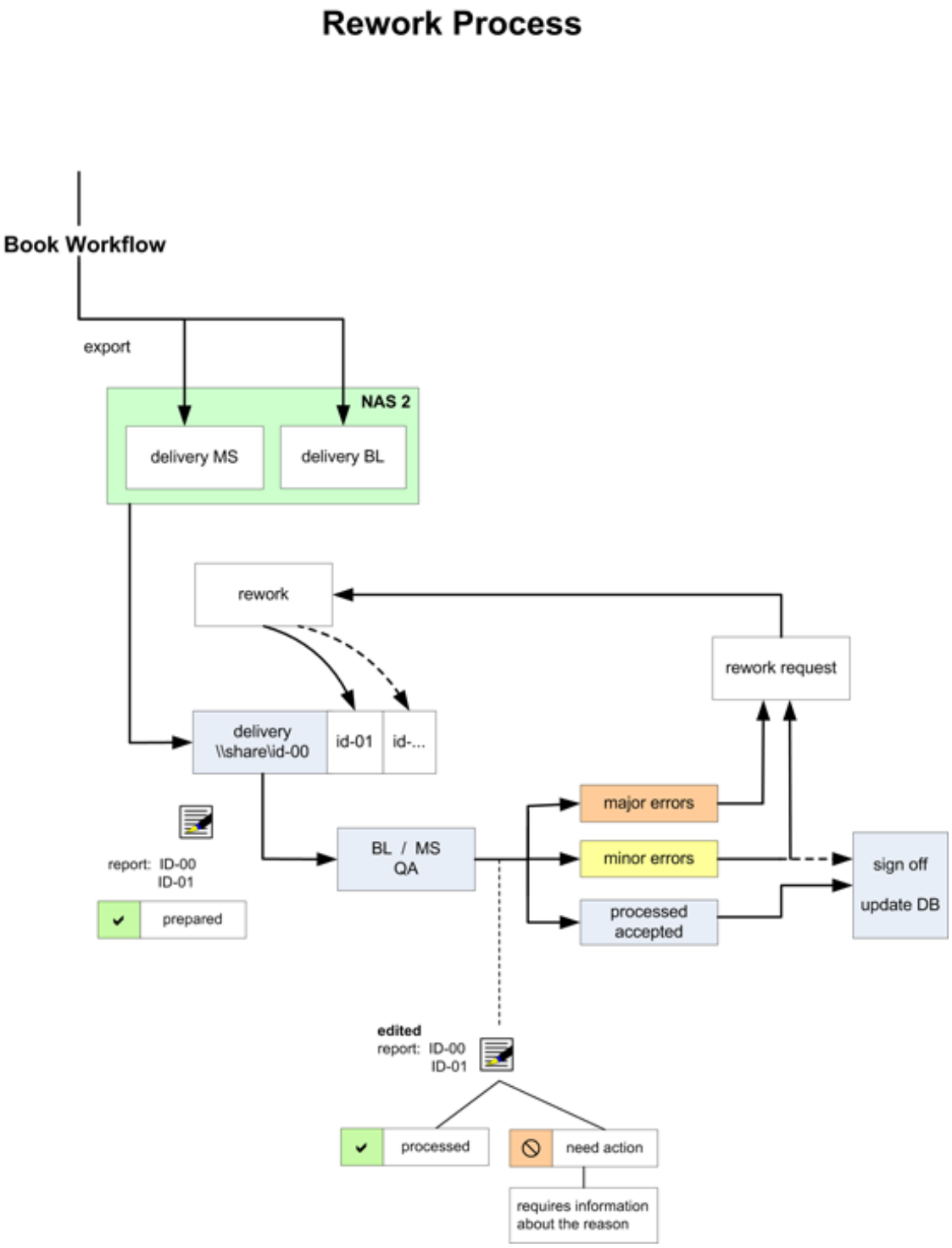

Image showing the QA with book digitisation workflow

Diagram supplied by Content Conversion Specialists GmBH: http://www.content-conversion.com/

Links and further reading

Implementing Quality Assurance for Digitisation; 2004; UKOLN: http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-27/html/ Retrieved 12.02.2010

Quality Assurance; 2008; JISC Digitial Media: http://www.jiscdigitalmedia.ac.uk/advice/creating/pdf/qassurance.pdf Retrieved12.02.2010

Delivery, Presentation and Dissemination

When delivering the content, lots of parameters need to be balanced: the type of collection, the prospective audience, the technical requirements, the storage space available etc.

Presentation can happen in image galleries, using thumbnails, link lists, monitors, or on removable media. It can include networks, monitors and printers and require the corresponding file format of the master image.

OCR output is usually received as text but can be marked up in an HTML file and/or exported to PDF for delivery.

In 2003, the Minerva network issued 10 cultural website quality principles: http://www.minervaeurope.org/immagini/postercwqp.pdf Retrieved 12.03.2011

User studies suggest that researchers expect fast retrieval, acceptable quality, and complete display of digital images. Several variables control access speed, including the file size, network connections and traffic, and the time to read the file from storage and to open it on the desktop. Additionally legibility and completeness often conflict and decision need to be taken to balance between them accordingly.

Questions to consider are:

- Who are the intended audience? What would be their preferred means of receiving this information?
- What functionalities should be provided to the user?
- Will users be allowed to edit the content (e.g. via tagging, or by manually correcting OCR results)?
- What size and format will the images be presented in?
- How should the metadata be presented (e.g. by embedding it in the digital object, or as separate explanatory matter)?
- Does the content necessitate any access restrictions (e.g. copyright)?

Data backup

Determine whether, when and in which intervals to refresh media and prepare for migration to final site. Questions to consider:

- What kind of backup system is needed?
- In what intervals does backup need to be done?
- Who will do it? Is material being outsourced?

See also the IMPACT Storage Estimator [https://www.surfgroepen.nl/sites/impactproject/oc/GA%20Annex%201%20Description%20of%20Work/OC2/OC2.1/IMPACT_Storage-Estimator_BSB_version3-2.xls] for the storage and cost implications of different types of files.