FRAGEN

QUING

Digitisation within Fragen

Manual

Fragen is a subproject of the QUING project
FP6 Integrated Project - 2006-2011

By Tilly Vriend

November 17, 2009

FRAGEN project
c/o Aletta, Institute for Women's History
Obiplein 4,
1094 RB Amsterdam
http://www.aletta.nu
+31 20 6651318
Fragen@aletta.nu

Content:

Introduction

Many libraries have experience with the digitisation of parts of their collections. Other organisations participating within Fragen probably don't share this experience. In order to understand the process of digitisation, it is explained below.

However: we strongly advise all Fragen partners to outsource this job to an external digitisation agency, who is familiar with international standards and has experience in this field / has been working for libraries!

1.What is digitisation?

The process of digitisation involves creating a reproduction of an existing physical object in the form of an electronic file which can be stored and accessed using a computer and viewed on a computer monitor. This can be done in-house or outsourced to an external digitisation agency. However the digitisation is done, it will be necessary to establish the appropriate file formats and resolution and to ensure that all items digitised meet the project's quality criteria.

How do you digitise a text?

This is done through the use of scanning and Optical Character Recognition (OCR) software. Criteria for what is appropriate for digitisation include:

- the suitability of the material for digitisation (paper condition; color; size; quality; content; etc.)
- how the collection should be prepared with respect to physical, organisational or intellectual considerations
- the extent to which text should be made machine-readable (i.e. choose some level of OCR or simply scan as image files)

Files should be created, stored and disseminated as simple raw images and converted to fully searchable text. Some materials may not be suited to OCR for instance hand-written comments.

2. Preparing paper for scanning

Correct preparation is important, when documents are scanned or OCR'd. It is important to work with the material following accepted conservation standards that respect the original media. Key stages of preparation include:

- assessment of material and its condition
- selection of material suitable for scanning
- removal of items such as staples and paperclips
- collation of original papers using archive quality paper clips and folders post scanning
- creation of inventory of materials (if absent)

3. Scanning

The original is scanned and saved in Tagged Image File Format (TIFF). TIFFS are single pages of the document. So after scanning one document will consist of many TIFFS. The original is retained at this stage for reference purposes during proofing and processing. After image scanning follows full capture of text. OCR packages are used to convert the scanned images to text.

Digitisation standards and requirements

File format	Definition	Ideal image
TIFF (Tagged Image File Format)	TIFF is an image file format used extensively for the storage of high-quality images. Appears as .tif.or .tiff	Best used for master images, is the defacto standard.
JPEG (Joint Photographic Experts Group)	JPEG is a compressible bit-map graphic format that can be saved in three different formats. .Appears as .jpeg, .jpg .jif or .jiff **The standard we use for FRAGEN is JPEG 7**	Everyday use/web presentation

The external digitisation agency will return the master files in principle as TIFF and/or JPEG (compression factor 10:12) 1:1 digitised in RGB (in color).

4. Resolution

This refers to the number of dots or pixels used per inch when undertaking the digital capture of each item. It is expressed in dpi (or dots per inch). In general the higher number of dots per inch an item is digitised at, the higher the quality of the resulting image. If it is necessary to view a high level of detail in the image, you will need to capture it at a higher dpi. Again, the higher the resolution, the larger the file size. For the Fragen project we recommend to scan texts at 300 dpi .

5.Searchable PDF

PDF and PDF/A — In April 2008, the United Kingdom's Digital Preservation Coalition (DPC), named Portable Document Format (PDF) as one of the best file formats to preserve electronic documents and ensure their survival for the future. This decision allows information officers to follow a standardized approach for preserving electronic documents. The DPC report suggests adopting PDF/Archival (PDF/A) for archiving electronic documents as the standard that will help preservation and retrieval in the future.

The use of Portable Document Format files (PDF) allows the look of the original papers to be preserved. The usefulness of a PDF version may be further extended by creating one which is text searchable. This can be a relatively quick process. For each document, all the constituent TIFFs are collated and converted to a PDF using 'Paper Capture' in Adobe Acrobat. This process is reasonably quick and efficient. Of course, the degree of searchability is much lower than fully OCR'ed text but it is secured at a much lower cost of time and labor.

NOTE: to make the contents of the text accessible and easy to use, we ask you to merge all pages (TIFFS or JPEGS) into one PDF (multipage) document. You might ask your external digitisation agency to make the PDF's for you!

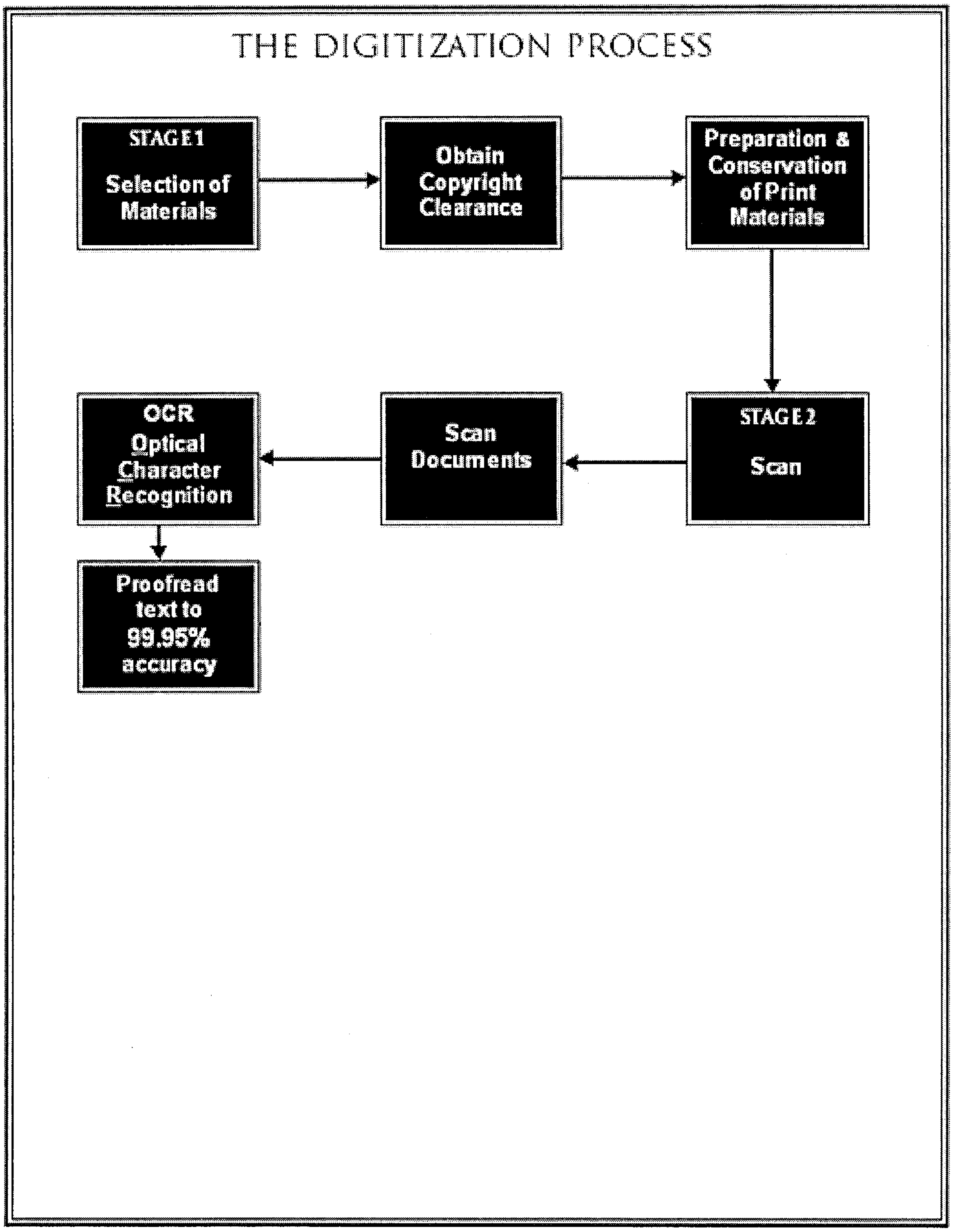
6.Naming the digital document

It is essential that each text will get a unique name/code which is needed to store the digital text in the database
This so called identifier will exist of 4 letters and 13 figures:

- 4 letters: the abbreviation of the name of the partner/organisation or the first 4 letters of the name of the partner/ organisation. In the case of Aletta, Institute for Women's History, this will be ALET
- The figures consist of the ISBN or ISSN of the text : for instance ISBN : 9781405149037 If there is no ISBN or ISSN on the text, contact Fragen.
- The unique identifier of the text will be: ALET-9781405149037. The pages will get the same number followed by 001.jpg -002.jpg.
- PDF's will be saved as ALET-9781405149037.pdf
- OCR's will be saved as ALET-9781405149037.txt
- In the database we would like to show the image of the cover of the document (frontpage) of the document
  So, most probably, the name of the cover of the text will be: ALET-9781405149037-001.jpg; (always check if this is correct!)

NOTE: please contact the Fragen management, if your organisation uses different schemes and/or when you have further questions.

Appendice: Flow-chart digitisation process page 5


THE DIGITIZATION PROCESS
STAGE 1 Selection of Materials
Obtain Copyright Clearance
Preparation & Conservation of Print Materials
STAGE 2 Scan
Scan Documents
OCR Optical Character Recognition
Proofread text to 99.95% accuracy

Appendice 1