DELIVERABLE

Project Acronym: thinkMOTION

Grant Agreement number: 250485

Project Title: Digital Mechanism and Gear Library goes Europeana

---

D.5.1 - Intermediate report on enhanced digitised and online available content accessible from Europeana

Revision: 1.2

Authors:
Erwin Christian Lovasz (University Politehnica of Timisoara)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	x
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
1.0	15.06.2011	Lovasz. E.-C.	UPT	Annual report
1.1	12.07.2011	Veit Henkel	IUT	Review
1.2	19.07.2011	Rike Brecht	IUT	Review

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1 Introduction

The objective of WP5 is to achieve a proper web-content representation in a high quality and quantity for a broad range of users by processing approved processing steps. The result of this work package is the online available content accessible for Europeana.

The tasks to carry out with WP5 are:

- Task 5.1: OCR (Optical Character Recognition) of textual content.
- Task 5.2: Quality improvement of digitised content
- Task 5.3: Import of digital existing content
- Task 5.4: Converting into web-compliant content
- Task 5.5: Integrating the content into the online portal of DMG-Lib accessible for Europeana.

WP5 performs activities which connect logically and coherently with tasks in other work packages.

2 Workflow and results

Different classes of items, such as documents, images, animations, CAD applications and so on, are processed in WP5. Figure 1 illustrates, for example, the steps needed to obtain an item available online starting with an analogue document.

Figure 1. General workflow for processing documents

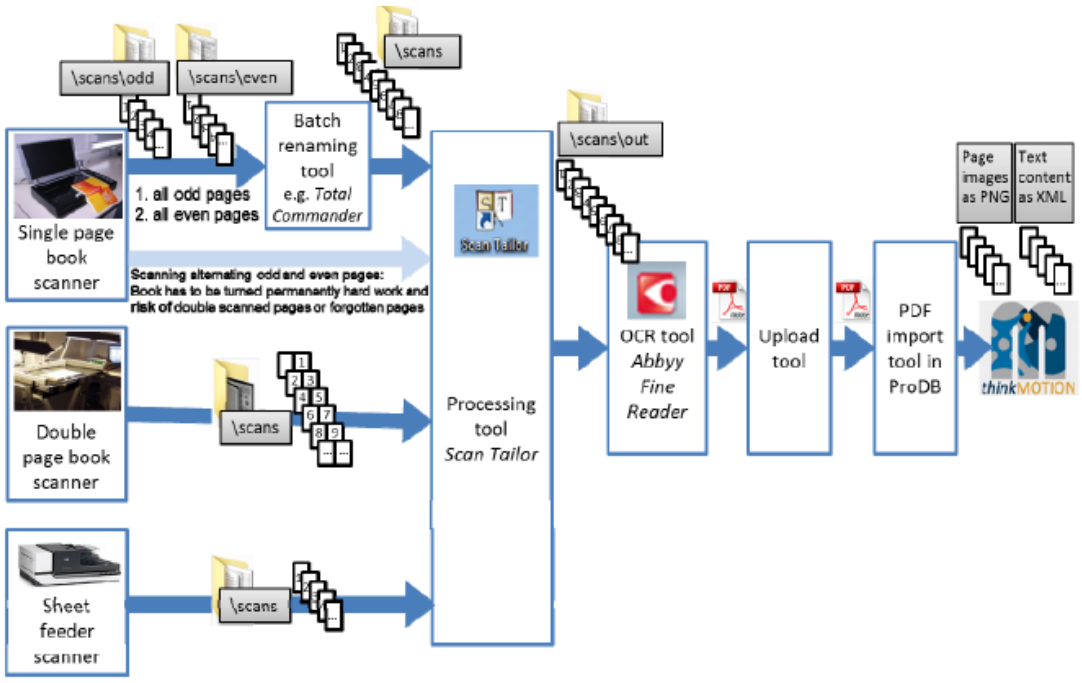


Figure 2. Operation of splitting pages of a book and automatic deskewing of a page

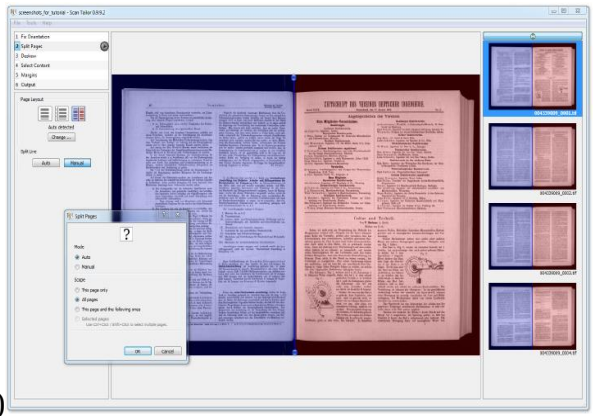
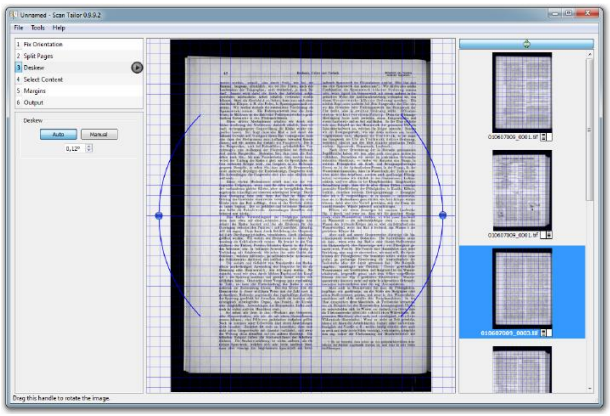
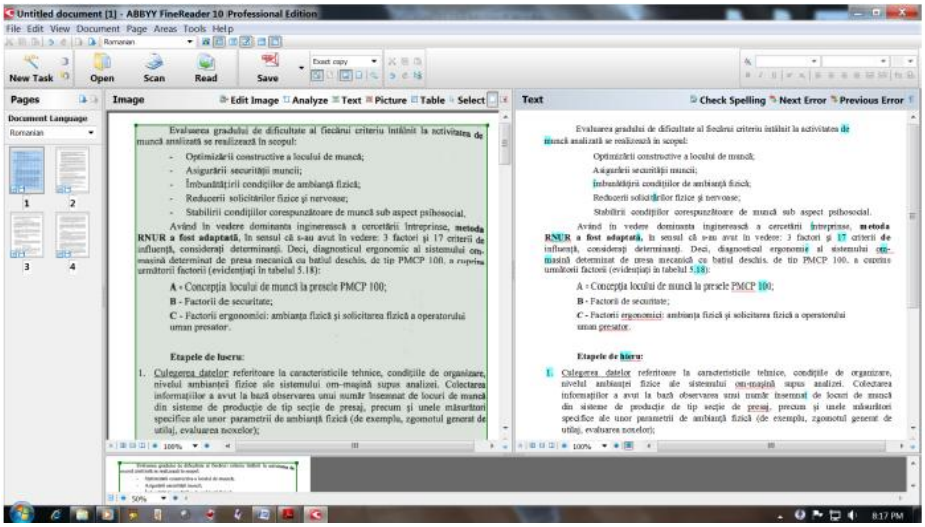
a)  b) 

Figure 3. Optical recognition of characters with ABBYY FineReader



A scanned analogue document results with unavoidable flaws (tilt of written content, non-uniform contrast etc.) and requires an improvement of quality. There are special software tools, such as ScanTailor, to perform actions such as: splitting the scanned double-pages of a book (Figure 2a), deskewing the content (Figure 2b), cutting the useless margins, resolution and colour mode setting etc.

If the image quality is set to the desired quality level, then the file is converted to a searchable PDF file. At this stage of processing, WP5 used professional software – ABBYY FineReader. Figure 3 shows the operating with FineReader.

Images and photos can be enhanced using specific software such as Adobe Photoshop or IrfanView.

Content published in the last few years often exists in digital form such as PDF or MS Word format. In this case the step of digitization is skipped and the content is imported into the internal data format for further processing steps.

The output of task 5.2 and task 5.3 must be converted into a web-compliant version. In this processing step the color depth and the resolution will be reduced and the content is stored in a web-compliant format. For textual content PNG file format is used. To allow a full text search and highlighting of found words in the online portal text reader, the full text must stored with information about the position of each word in XML-files. The digitized pictures and slides are stored in JPEG file format. The taken image sequences of the physical models are converted into video files (MPEG format) and into an interactive animation file format playable by a Java applet inside the DMG-Lib portal.

The last step in the workflow is uploading and integrating the digitized and enhanced content into the DMG-Lib portal. After this step the content is online and accessible from the DMG-Lib portal and also from the Europeana portal.

3 Results

The work progress and detailed achievements of all partners in the project are summarized in Table 1.

Table 1. Work progress in WP5 considering tasks and item classes

Task 5.1 - OCR (Optical Character Recognition) of textual content				
Pages in documents				
29,816				
Task 5.2 - Quality improvement of digitised content				
Pages in documents	Images	Movies		
37,612	9,352	356		
Task 5.3 - Import of digital existing content				
Pages in documents	Images	Movies	Animations	CAx models
19,498	12,280	32	65	2117
Task 5.4 - Converting into web-compliant content				
Pages in documents	Images	Movies	Animations	CAx models
48,192	6,395	79	55	2117

The following notices describe the way teams accomplished the work:

- all teams used perform activities of WP5 with their own project staff;
- fulfilment of tasks required the use of a wide range of professional software tools, such as ABBYY FineReader v.10, Scan Tailor v.9.9.2, Irfan View v.10, Nitro PDF v.6.2, Adobe Acrobat Pro v.10.4, CATIA, ADAMS, Camstudio, different scanning software and ProDB;
- the workshop at Ilmenau in February 2011 gathered members of all teams. The discusions lead to the improvement of the workflow. The knowledge got during the training was spread among the local staff of each team through subsequent local trainings. One of the most important result of the workshop at Ilmenau was the development of a tutorial for digitizing and processing paper based documents (see Annex I).

Significant results may be highlighted, as follows:

- a very large amount of items are processed up to different stages of work. On a few amount of the content already existing in digital form are processed in WP5 during

the first project year. Most items produced in the first project year needed performing of all steps in the workflow. Quality of digitized content was pursued as well as the scientific value of items. CAD model were generated using advanced complex knowledge;

- the workflow was improved so that all teams can work at the same level of quality and efficiency;
- a valuable tutorial regarding the tasks in WP5 was developed (see Annex I).

During the first year, the partners in the WP5 got the expertise required for high-quality, efficient activity for the next two years.

Annex I

Tutorial Workflow for Digitizing Paper Based Documents for the thinkMOTION Project
(see D4.1 - Intermediate report on the work package WP4 “Digitising heterogeneous input content” of the project thinkMOTION)