DELIVERABLE

Project Acronym: thinkMOTION

Grant Agreement number: 250485

Project Title: Digital Mechanism and Gear Library goes Europeana

---

D5.2 - Intermediate report on enhanced digitised and online available content accessible from Europeana

Revision: 1.1

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	01.06.2012	E.-C. Lovasz	UPT	Annual report
1.1	30.06.2012	S. Falke	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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Table of Contents

1 Introduction

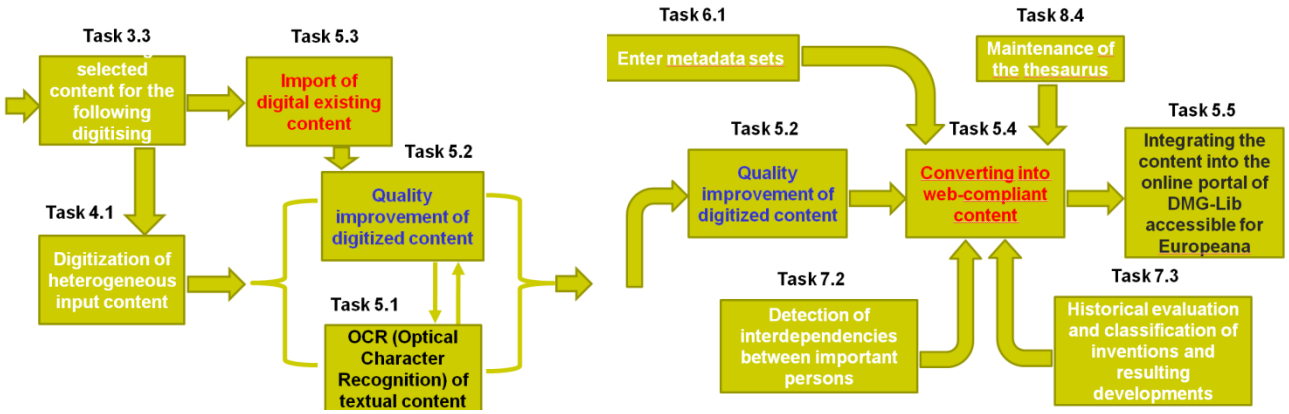
The objective of WP5 is to achieve a proper web-content representation in different levels of qualities and to make the content acceptable to a broad range of users by processing an approved workflow. The result of this work package is the online available content accessible for Europeana.

The tasks to carry out with WP5 are:

- Task 5.1: OCR (Optical Character Recognition) of textual content.
- Task 5.2: Quality improvement of digitized 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. The dependencies of WP5 with other WPs are illustrated in Figure 1.

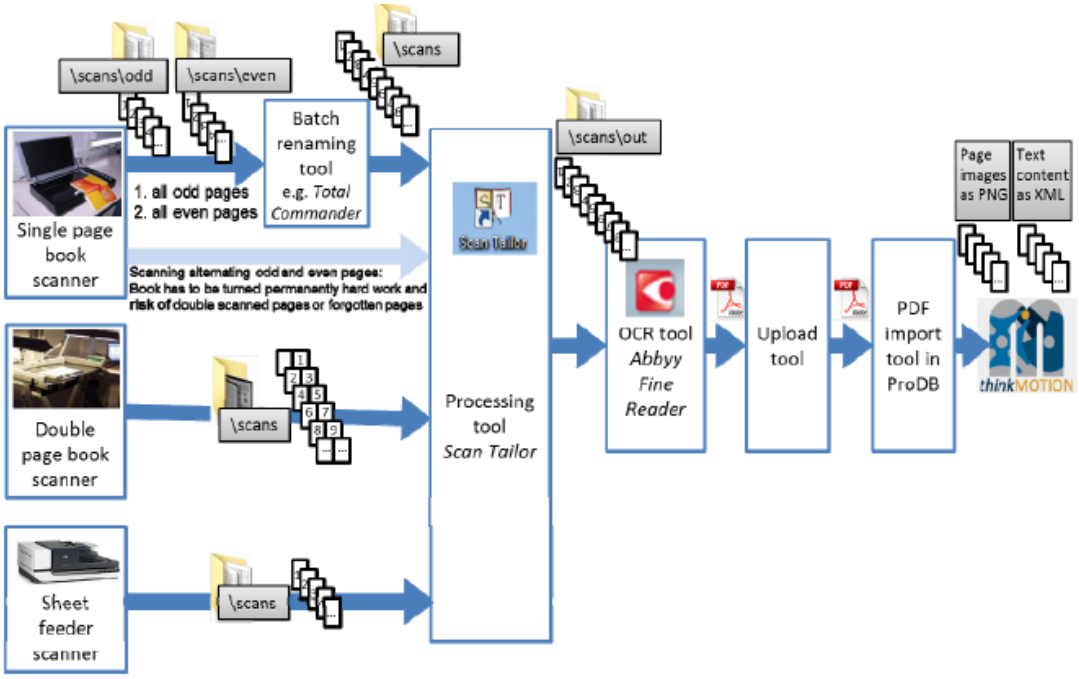
Figure 1. Dependencies of WP5 with other WPs



2 Workflow and results

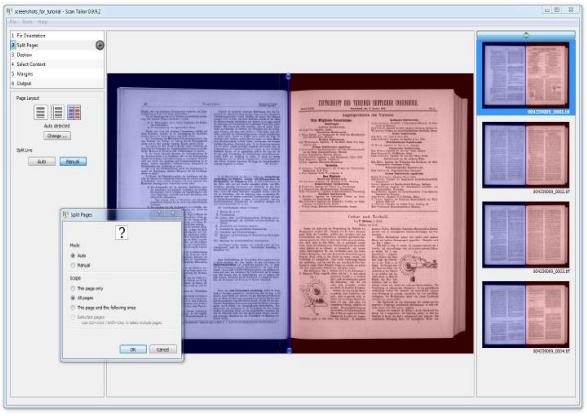
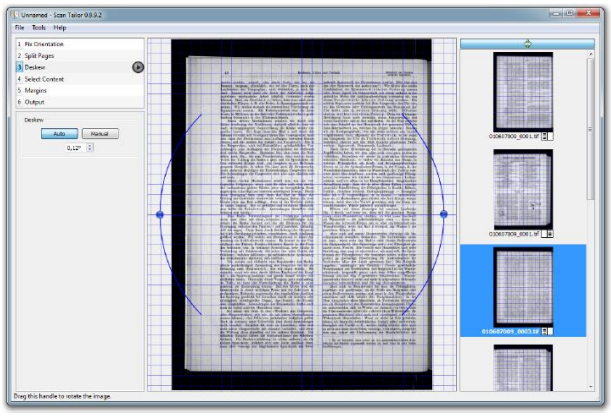
Different classes of items, such as documents, images, animations, CAD applications and so on, are processed in WP5. Figure 2 illustrates, for example, the steps needed to obtain an item available online starting with an analogue document.

Figure 2. General workflow for processing documents



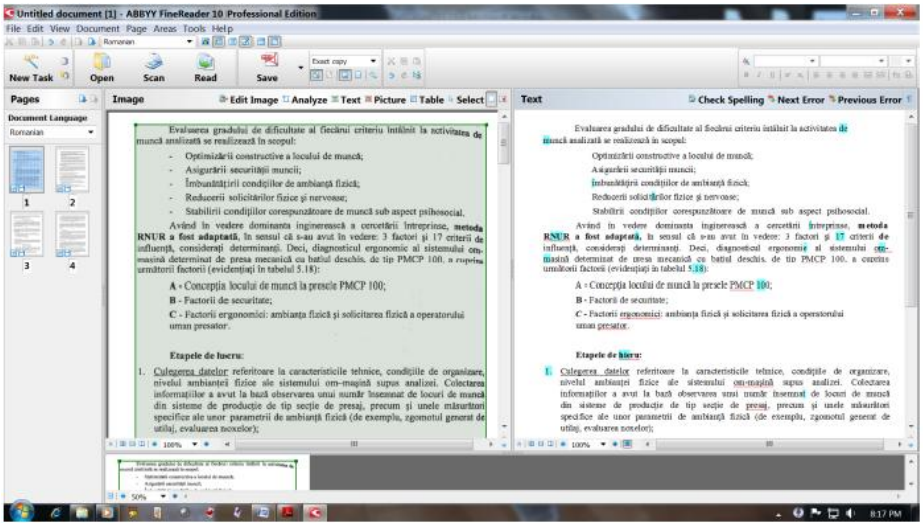
A scanned analogue document results with unavoidable flaws (tilt of written content, non-uniform contrast etc.) and requires an improvement of quality. There are special software tools, such as ScanTailor, to perform actions such as: splitting the scanned double-pages of a book (Figure 3a), de-skewing the content (Figure 3b), cutting the useless margins, resolution and colour mode setting etc.

Figure 3. Operation of splitting pages of a book and automatic de-skewing of a page

a)  b) 

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 4 shows the operating with FineReader.

Figure 4. Optical recognition of characters with ABBYY 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 colour 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 be 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

During the second year of the project, the activity in WP5 was much more efficient than in the first year and resulted in production of a large amount of items available online. The work progress and detailed achievements of all partners within the tasks of WP5 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					
306,569					
Task 5.2 - Quality improvement of digitized content					
Pages in documents	Images	Movies	Physical models	Animations	CAx models
315,073	9,987	1,283	890	569	252
Task 5.3 - Import of digital existing content					
Pages in documents	Images	Movies	Physical models	Animations	CAx models
36,032	4,826	235	13	0	0
Task 5.4 - Converting into web-compliant content					
Pages in documents	Images	Movies	Physical models	Animations	CAx models
223,836	9,916	500	900	569	19
Task 5.5 - Integrating the content into the online portal of DMG-Lib accessible for Europeana					
Pages in documents	Images	Movies	Physical models	Animations	CAx models
165,364	11,870	500	900	569	19

The following notices describe the way teams accomplished the work:

- Most teams used only persons in the project to perform activities of WP5
- Fulfillment 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

Significant results may be highlighted, as follows:

- A very large amount of items are processed up to different stages of work; very few content exists in digital form; most items need performing of all steps in the workflow
- Quality of digitized content was pursued as well as the scientific value of items
- The workflow proved to be correctly conceived so that efficiency of work increased compared to previous year
- CAD models were processed using advanced complex knowledge and appropriate software
- The partners approached various classes of items
- A series of valuable and rare historical books and journals from the beginning of the twentieth century were processed
- Complex software able to synthetize images and produce digital models was implemented.

During the second year, the partners in WP5 used their expertise from the first year to fulfill related activities efficiently and with high-quality. A large amount of content is still waiting for signatures of the authors before posting online and the capacity of data transfer is limited by the serves. The partners are able to solve these minor deviations within the last year and to increase efficiency of item production again.