



DELIVERABLE

Project Acronym: thinkMOTION

Grant Agreement number: 250485

Project Title: Digital Mechanism and Gear Library goes Europeana

D4.1 - Intermediate report on digitised input content

Revision: 1.1

Authors:
Veit Henkel (Ilmenau University of Technology)

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	29.06.11	Veit Henkel	IUT	
1.1	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This deliverable contains original unpublished work except where clearly indicated otherwise.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.....	4
2	Workflow of the work package WP4 - Digitising heterogeneous input content	4
2.1	The aim of the work package WP4 and dependencies on other WPs.....	4
2.2	Workflow of digitising paper based documents	4
2.3	Scanner hardware used in the thinkMOTION project.....	8
2.4	Scanner software used in the thinkMOTION project	12
2.5	Scanning parameters and file format for saving the raw scans.....	13
2.6	File naming and saving path.....	14
2.7	Additional file handling for special pages like cover or fold-out pages	15
2.8	Additional file handling for „Single page book scanners“	15
2.9	Quality Check Up	17
3	Results of the work package WP4 within the thinkMOTION project	17
4	Outlook for the next project year	18
	Annex I	19

1 Introduction

In June 2010 partners from six European universities started the project thinkMOTION with the main objective of providing content from the field of motion systems via the Europeana online portal.

The located content, which is proved to be relevant, will be digitized, processed and presented in the multilingual interactive online portal DMG-Lib (www.dmg-lib.org) accessible also via Europeana and by this way also accessible for a wide range of European user groups like interested laymen, engineers, scientists, lecturers and students. The provided interactive material leads to a deeper understanding and motivates to learn more about the scientific and technical background in a European society of lifelong learning. Very different types of content like textual sources, physical models and drawings will be collected by partners in several European countries. In the library, new ways of content representation, information retrieval (supported by a multilingual semantic network) and cross-linking are supported.

The thinkMOTION project is divided into ten work packages, which depend partly from each other.

Work package 4 (abbreviation: WP4) with the title “Digitising heterogeneous input content” has the aim to digitize heterogeneous content sources using different special work places in a high quality and quantity. This paper is an intermediate report about the work in WP4, which was done in the first project year.

2 Workflow of the work package WP4 - Digitising heterogeneous input content

2.1 The aim of the work package WP4 and dependencies on other WPs

The aim of the work package WP4 is to digitize heterogeneous content sources called input content in the following. For content which already exists in digital form (e.g. proceedings of conferences of the last few years, which can be provided in PDF or MS Word file format) the workflow in WP4 can be skipped. Such content will be processed in WP5 – “Processing of digitised content and integration into DMG-Lib”.

In the first project year the focus was on the digitizing of paper based documents such as books, journals, journal articles, patents, theses, proceedings, technical drawings, drawings from printed documents, teaching material. Additionally, some of the project partners have taken photos of machines, devices or portraits or videos of physical demonstration models or have scanned slides.

The work in WP4 (Digitising heterogeneous input content) is depending on other work packages and tasks. WP4 receives input from WP3 (Locating and providing relevant sources and clarification of rights of use). In WP3 the located and selected content will be registered in the DMG-Lib database, the Intellectual Property Rights (IPR) for these input contents will be clarified and necessary rights of use will be obtained. The work in WP3 provides the selected content for the digitising process in WP4. WP4 delivers input for WP5 (Processing of digitised content and integration into DMG-Lib). In WP5, the scanned documents will be OCR processed, quality improved, converted into a web-compliant file format and uploaded into the DMG-Lib online portal accessible for Europeana.

2.2 Workflow of digitising paper based documents

In the thinkMOTION project a MySQL Database, running on a SuSE Linux Apache server is used for storing the content data, the meta data and the process data, which describes and register the data flow and all important workflow steps. This database is called thinkMOTION *Production Data Base*, in the following abbreviate with ProDB.

Figure 1 shows the workflow necessary to scan, process, and upload a document. This workflow involves two computers: the ProDB server and local computers that are connected to the scanner hardware at the partner’s workstations.

The four main steps in the workflow are:

1. **Reserving** the document on the ProDB server and creating a folder structure on the local computer: This will lock the document and shows other users of the ProDB that they should not scan it. This prevents redundant work by scanning the same content at various partners. The folder structure will be created on the local scanner workstation computer and will hold the files for one document (scanned and processed content).
2. **Scanning:** Depending on the type of the used scanner hardware, the scanned documents have to be saved into a specific folder of a ProDB-created folder structure by using a predefined file name on the local computer.
3. **Processing:** To improve the visual quality of your scanned images and to make the text of the documents searchable, some processing steps are necessary. This work is a task of WP5.
4. **Uploading:** As soon as scanning and processing is done on the local computer and the folder structure contains all necessary files, it can be uploaded to the ProDB server. There it will be archived, imported and linked to the corresponding database entries. This work is also a task in WP5.

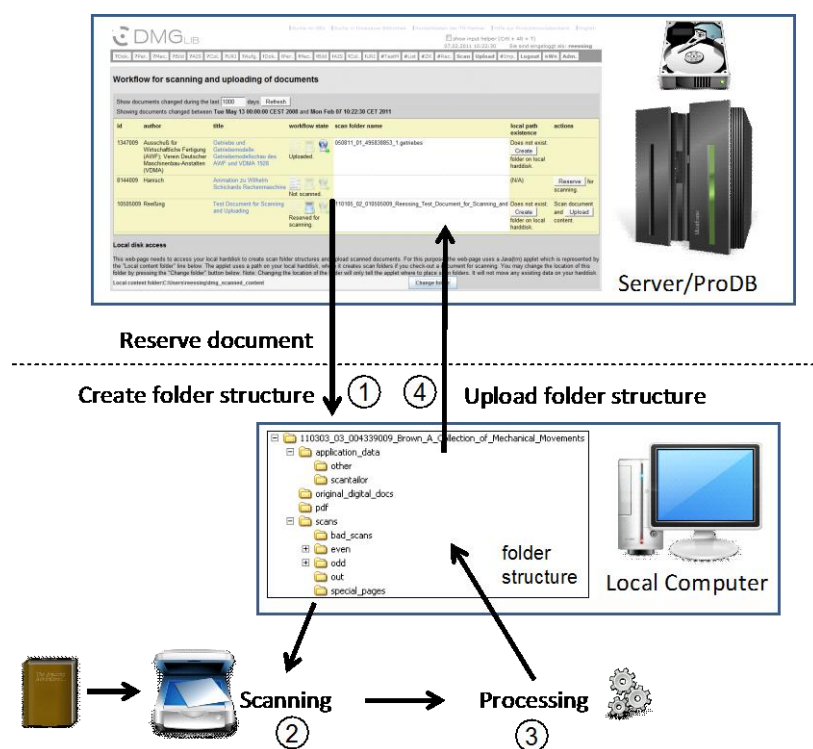


Figure 1 - General workflow for reserving and uploading documents – contains tasks of WP4 and WP5

In the ProDB the “Scan Workflow Overview” page displays a table of documents that you have edited recently (Figure 2, Label 1). Each line of the table contains the DMGLib ID, the author(s), the title and the workflow state of one document (Figure 2, Label 2). The state can be “not scanned”, “reserved for scanning” or “uploaded”.

The not-scanned-state means that the document has not been scanned or reserved for scanning. The reserved-for-scanning-state shows the operator, that the document has been reserved and will be scanned and processed soon. The third state is named uploaded and means that the document has been scanned processed and uploaded. In this case, the workflow steps of digitizing and quality improvement (WP4 and WP5) are finished.

Other columns of the ProDB-workflow-table are the scan folder name the server has assigned to this document and a column (local path existence) that indicates if this scan folder exists on your local hard drive in the content folder (see Configuration section above). If a scan folder name has been assigned, but does not exist on the local computer, the column contains a button to create the folder on the local hard disk.

The last column lists available actions that depend on the state of each document. These actions can make a reservation for scanning or can start the uploading process of the content (raw scan data and processed data). Actions can be reserve for scanning and upload content. The Reserve-for-scanning-action reserves the document entry for scanning and processing. The document entry is then locked. This state will be stay until you have uploaded the scanned and processed content. The Upload-content-action starts the uploading process for the raw data and the processed content from the local scanning computer to the server.

Scan Workflow Overview

Show documents changed during the last 31 days. Refresh

Showing documents changed between Mon Jan 31 00:00:00 CET 2011 and Thu Mar 03 11:47:00 CET 2011

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	Not scanned.		(N/A)	Reserve for scanning.
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	Not scanned.		(N/A)	Reserve for scanning.
10505009	Reeßing Test Document for Scanning and Uploading	Not scanned.		(N/A)	Reserve for scanning.

Local disk access

This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk.

Local content folder: C:\scans

Change folder

Figure 2 - Workflow state and reserving documents

By reserving a document for scanning in the ProDB, a folder structure (Figure 3) will be created on the local computer in the chosen local content folder path.

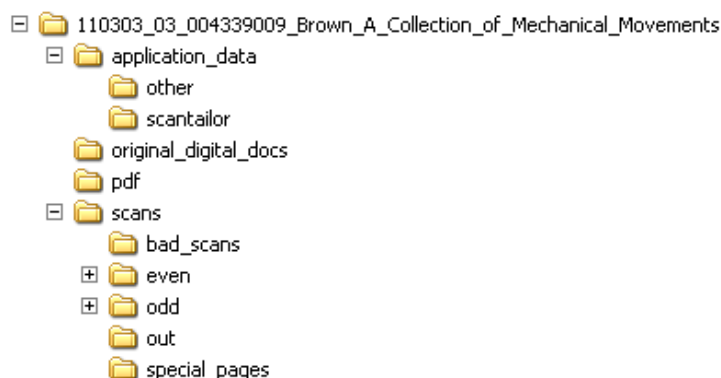


Figure 3 - Example for the predefined folder structure created on the local hard disk

The names of the generated folders means:

Main folder name: yymmdd_##_xxxxxxxxx_authorsname_begin_of_title

yymmdd	Date of starting process (2 digits for year, month and day)
##	Serial number of source of the day (start index = 01 for each day)
xxxxxxxxx	9-digit DMG-ID (creating by ProDB)
authorsname	Name of the main author or editor
begin_of_title	The first significant words of the title

Subfolders are used for raw scanned data, processed data or processing project data. Table 1 shows the meaning and the usage of the folders.

Table 1 – Meaning and intention of the folders and subfolders generated by the ProDB

(“~\” means the path, which is used for the download of the folder structure from the ProDB on the local computer)

Subfolder name	Description what folders contain
~\application data	Project files of used tools such as the quality improvement tool <i>ScanTailor</i>
~\pdf	Output of the OCR tool <i>ABBYY FineReader</i> as PDF/A-File Not for other PDF documents such as described in ~\original_digital_docs section
~\original_digital_docs	Original files, which already exist as files and do not need a scanning process such as MS Word or PDF documents
~\scans	Scanned and unchanged images, the so-called raw scans, as uncompressed or lossless compressed files. Use this folder for scans from a double page book scanner, from flatbed scanner or from a sheet feeder scanner. Should be used also for fold-out pages scanned in the fold-in state.
~\scans\bad_scans	Scanned images, which are incorrect and rescanned again should be moved into this folder and do not be deleted.
~\scans\even	See ~\scans This folder should be used for scans of the even pages (2,4,6,8, ...) if a single page book scanner is used.
~\scans\even\bad_scans	See ~\scans\bad_scans but for scans of even pages
~\scans\odd	See ~\scans But use this folder for scans of the odd pages (1,3,5,7, ...) if you use a single page book scanner.
~\scans\odd\bad_scans	See ~\scans\bad_scans but for scans of odd pages
~\scans\out	Processed output images of the quality improvement tool <i>ScanTailor</i>
~\scans\special_pages	Scanned images of special pages such as fold-out pages scanned in the fold-out state, cover pages if in the original state, etc. (Figure 4) Images in this folder are the unchanged raw scans as uncompressed or lossless compressed files.

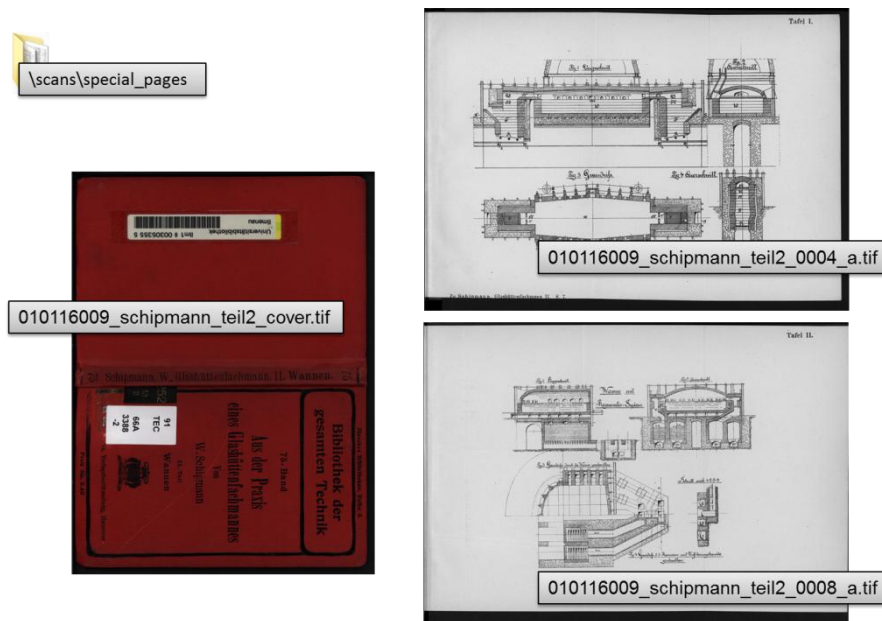


Figure 4 – Special pages such as a cover page and two fold-out pages scanned in the fold-out state

After reserving the document in the ProDB described above, the document must be scanned. The aim of this workflow step is to make image files from paper sources (books, articles, etc.) and to store them in the mentioned folder structure according to a given name scheme. Because of further processing steps all pages have to be scanned, also the blank pages. Cover pages should be scanned if they are in the original state and not blank (Figure 4, left). So called “Fold-out pages” have to be scanned within the fold-in state in the normal scan sequence. After finishing the process by scanning the last page, the fold-out pages have to be scanned in the fold-out state again (Figure 4, right).

2.3 Scanner hardware used in the thinkMOTION project

Generally, three different types of scanners concerning the types of the output files are used in the project:

- Double page book scanner (delivers a sequence of images containing two pages in each file)
- Single page book scanner (delivers a sequence of images containing single pages, only one type of pages, the even or the odd pages are in an upright position – explanation in chapter 2.8 Additional file handling for „Single page book scanners“)
- Sheet feeder scanner (delivers a sequence of single pages in an upright position)



Figure 5: Zeutschel OS 4000, book scanner with a special book-cradle

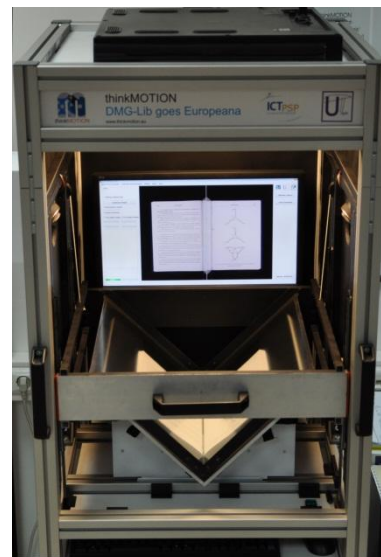
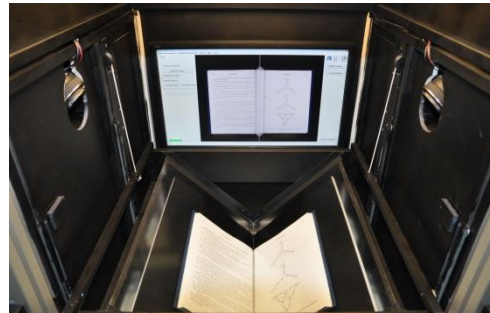
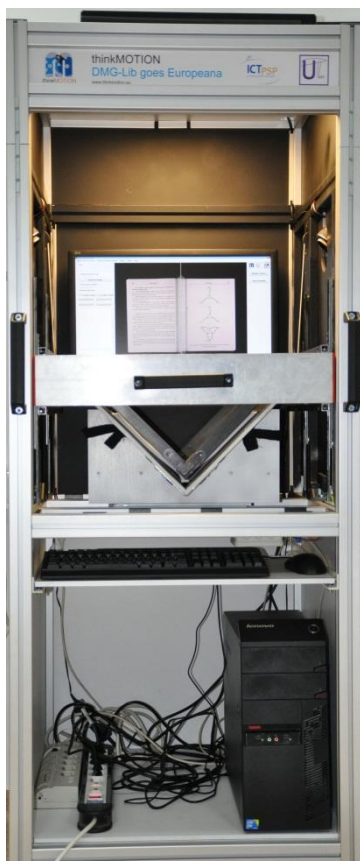


Figure 6: Self-made V-shape book scanner “tm-books”, developed by the Romanian partner



Figure 7: Book scanner Zeutschel OS 12000HQ



Figure 8: Plustek OpticBook A300, low-cost book scanner with a special book-edge

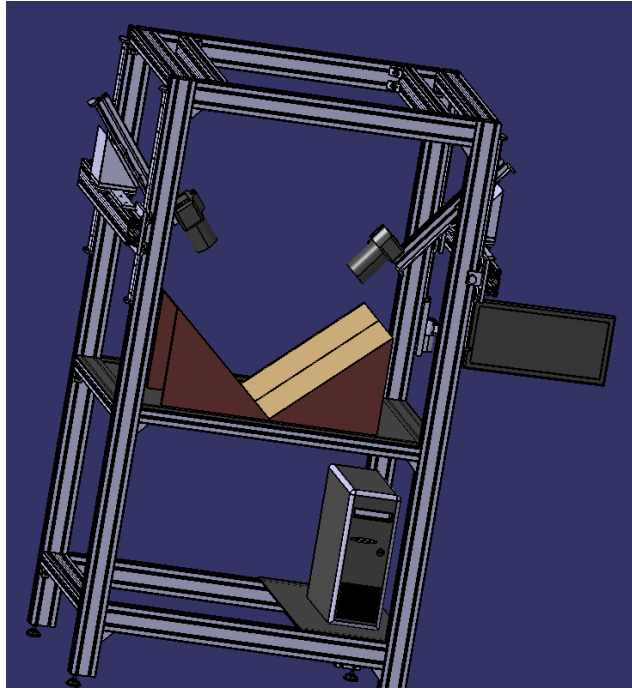


Figure 9: Self-made V-shape book scanner based on double EOS600D (18MPix each) developed by the French partner. Prototype will be tested in June 2011.



Figure 10: Epson GT-20000 with auto document feeder for loose leave documents

Figure 11 shows the different types of the output format of the raw images depending on the type of the used scanner. That means that different steps in the further workflow are necessary.


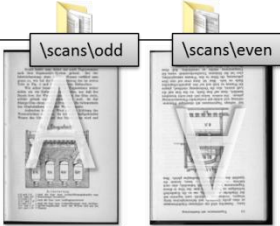




Type of scanner	Scanner (software) output
 Single page book scanner	<p>Odd and even pages in separate directories; even pages are upside-down</p> <p>(Alternated scanning of odd and even pages in one directory: Book has to be turned permanently. It is a hard work and a risk of double scanned pages or forgotten pages.)</p> 
 Double page scanner	<p>Two pages in each file</p> 
 Sheet feeder scanner	<p>One page in each file</p> 

Figure 11: Output of the scanner software depending on the type of used scanner hardware

2.4 Scanner software used in the thinkMOTION project

For each scanner an own scanner software is delivered or offered. Usually these software tools work probably and provide the main functionality of the hardware. Most of the project partners use the software, which was delivered with the scanner. However, in some cases the delivered scanner software does not work properly. E.g., the hardware of the book scanner *Plustek OpticBook A300* works very good and fast but the delivered scanner software causes a small but important problem. In the *Scan-to-file-mode* different picture file formats can be chosen, such as *TIFF-uncompressed* without any parameters and *TIFF-compressed* with some compression parameters – but both settings save the scan files in a JPEG-compressed TIFF-file with a loss of quality. However, it is a fundamental fixing in the thinkMOTION project to save the raw scan data in an uncompressed or lossless compressed file format.

Alternatively, we used a picture viewer with a batch scanning function via TWAIN interface, such as *IrfanView* (www.irfanview.com) and the latest vendor's scanner driver. But now there comes up a new problem. The device buttons for starting the scanning process at the *Plustek OpticBook A300* Scanner cannot be used while scanning via a third-party software. Therefore, the scanning process has to trigger with the mouse button or the PC keyboard. It is an uncomfortable handling because the operator has to press the book against the scanner's book edge, normally with two hands, in order to get a good scanning result and at the same time the operator has to start scanning process by the mouse button or the keyboard.

A well-proven solution could be found in the project by using a foot switch (Figure 12) in addition to mouse and keyboard. Therefore, the operator has both hands on the book and can trigger the scanning process with his or her foot. It is even easier than the usage of the original device buttons.



Figure 12: Foot switch for triggering the scanning process

2.5 Scanning parameters and file format for saving the raw scans

The following parameters are defined for scanning paper-based documents within the thinkMOTION project:

- **Colour Mode:** Scan colour mode depends on your type of document
 - All pages have to be scanned in colour mode with 24 bit, “True Colour”, 16,7 Million colours for documents with more than about 5% coloured pages.
 - All pages have to be scanned in grey scale with 8 bit, 256 grey scales, for all black and white or grey shaded documents.
 - All pages have to be scanned with grey scale and additionally all colour pages in colour mode if the document contains less than 5% colour pages.
 - 2 bit, bitmap or black and white scan modes are not allowed for scanning in this project.
- **Resolution:** Generally, the documents have to be scanned with a resolution of at least 300 dpi. Documents with small lines and small fonts should be scanned with 400 dpi. The resolution within a document should not be changed, because it causes problems in later steps.
- **Quality improvement tools of scanner software:** In normal case, it is not allowed in this project to use any kind of filters and image corrections functions such as brightness, contrast, sharpness etc. during the scanning process. They should be switched off. The quality improvement process takes place in WP5. The raw scan data have to be saved unchanged.
- **Scanning area:** To reduce the file size the scanning area can be reduced and be adjusted to the page size. However, it is necessary to check that no content disappears. Reducing the scanning area can speed up the scanning process significantly. The scanning area (width and height) for the same size of pages within a document should be equal.

The file size of the scanned images can reach approximately between 5 and 70 MB depending on the resolution, the colour mode and the page size. Therefore, an adequate disk space at the local scanning computer is essentially. Recommended is a free disk space of at least 1 Terabyte.

Scan files must be saved in a TIFF uncompressed file format (lossless compression as TIFF-LZW is possible to reduce the file size but not recommended). The lossless TIFF-ZIP compression cannot be used in this project because the software ScanTailor, which is used for further processing steps in WP5, cannot import this file format. Any lossy compressions such as JPEG are not allowed for scanning paper-based documents in the thinkMOTION Project.

2.6 File naming and saving path

The scanned files have to be saved in the folders, which were generated by the ProDB website automatically and described in Table 1. The scanned files should be named according to the following file name scheme. Usually the scanner software can automatically generate file name suffixes (running numbers). The operator has to initialize the start index, the number of digits and the increment value.

File naming: xxxxxxxxx_####.tif

xxxxxxx 9-digit DMG-ID (see DMG-Lib – ProDB, the same like in folder name)

4-digit serial number with the following parameters:

- start index = 0001;
- increment for the next single or double page image = 1;
- index for scanning of fold-out pages in the fold-out state = “page index of the fold-out page scanned in the fold-in state”+[a-z],

Example:

- image of the fold-out page scanned in the fold-in state:
001547009_0067.tif,
- image of the fold-out page scanned in the fold-out state
001547009_0067a.tif (Figure 4), and if it is too large for the scanner hardware, it have to be scanned into multiple images and b, c, d ... have to be attached to the file name such as 001547009_0067b.tif ...

Depending on the used type of the scanner hardware and the type of scanned pages different folders (Figure 3) should be used for saving the scanned images. The scanned images have to be saved into the folder “~/scans” or in the corresponding subfolder generated by the “Reserving a Document for Scanning and Processing” – process described above. Figure 14 shows the general workflow for processing documents and especially the different handling for the output of different scanners.

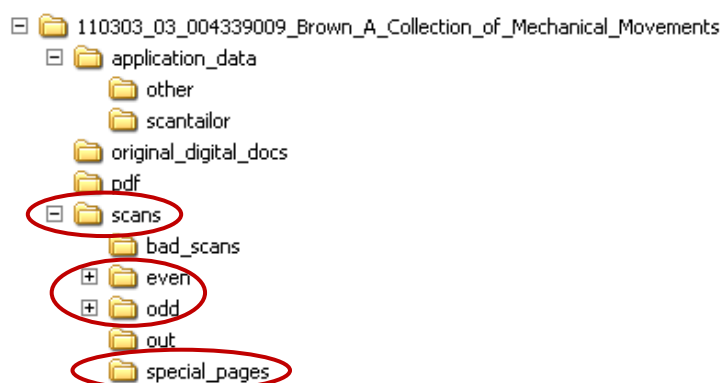


Figure 13 – Folders, which should be used for saving scanned images, depending on the type of scanner and the types of pages

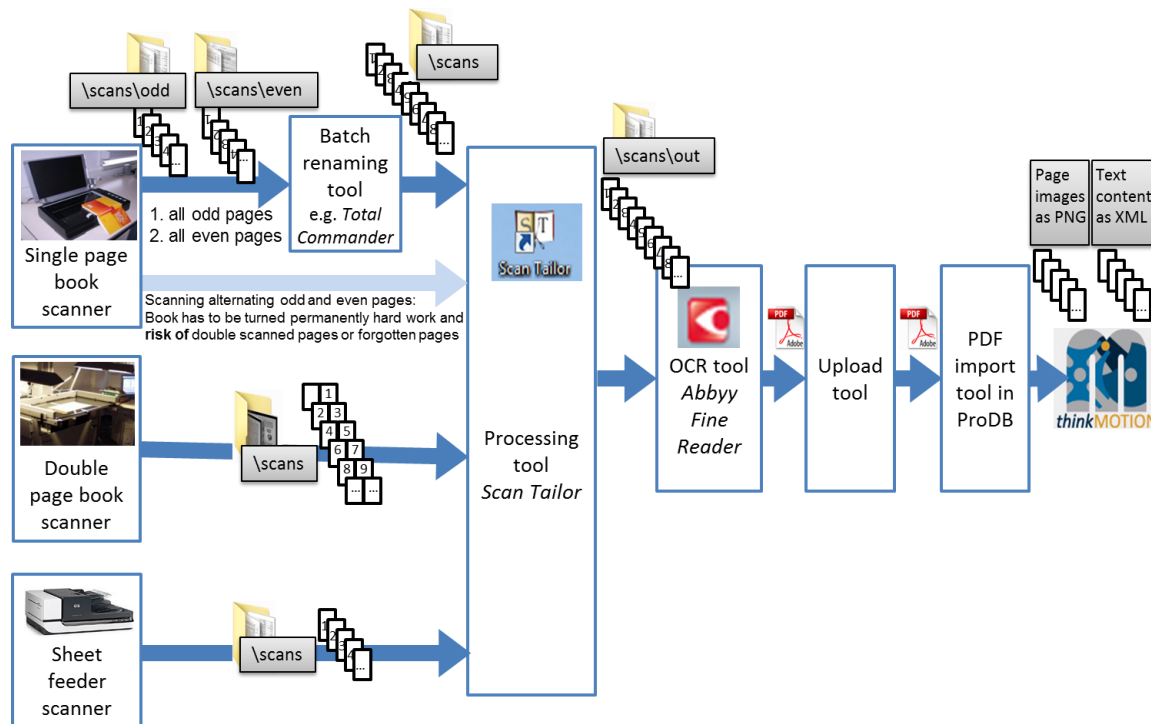


Figure 14 – General workflow for processing documents, folders for saving scanned images depending on the type of scanner hardware as well as the flow of generated documents

2.7 Additional file handling for special pages like cover or fold-out pages

Cover pages, if they are not blank, and fold-out pages in the fold-out state should be scanned into the folder “~\special_pages”. These pages have to be processed separately from the normal content pages. After finishing the quality improvement workflow steps in WP5 and before starting the OCR process, these special pages have to be merged with the normal scanning sequence.

The images of the fold-out pages in the fold-out state should be inserted after the fold-out pages scanned in the fold-in state in the normal scanning sequence. So the user can see, while page turning in the online portal, these pages first in the fold-in state and then in the fold-out state, just as in the original paper document. Before inserting these pages, the correct file naming as defined above has to be rechecked, so that the pages are in the correct alphabetic order.

2.8 Additional file handling for „Single page book scanners“

For single page book scanners, such as the Plustek OpticBook A300, scanning of odd and even pages alternately (means scanning following the page numbers) leads to errors such as twice scanned pages or forgotten and not scanned pages, because the book has to be turned after each scan. Another problem is the additional physical work load for the operator’s wrist and arm by turning the book permanently. For this reason, it is recommended to scan all odd pages first and then all even pages in two different folders.

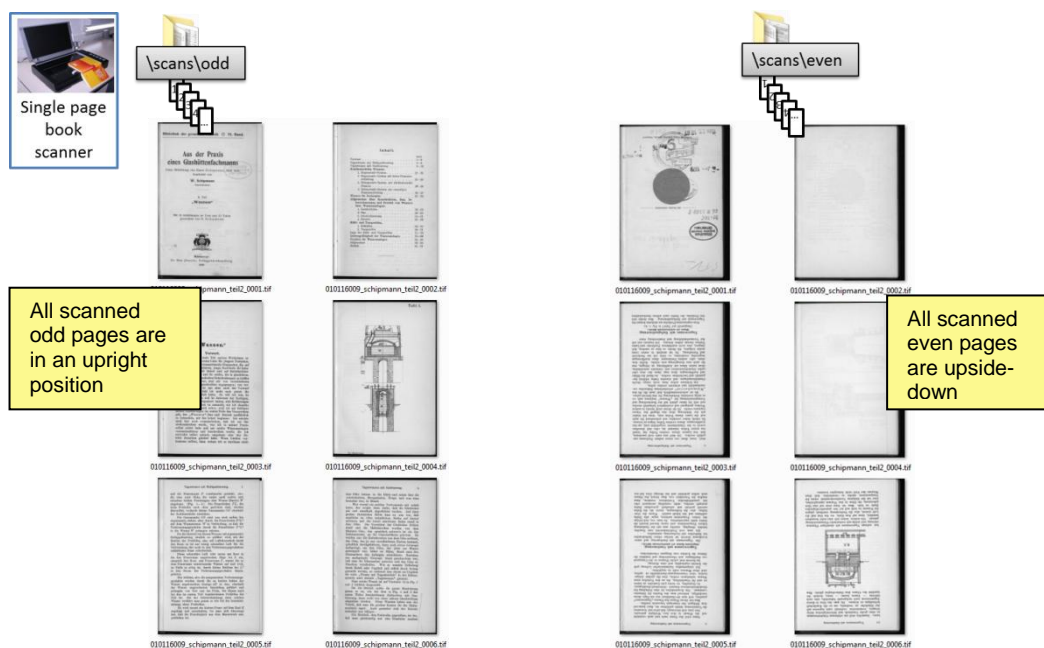


Figure 15 – Example of scanned pages in the odd and the even folder

The scanned odd pages are in an upright position and all even pages are upside-down (Figure 15) and have to be turned in a further workflow step with the tool *ScanTailor* in WP5 automatically.

After finishing the scanning process with a single page book scanner both scanning sequences have to be merged to one sequence in the folder “~\scans”. A simple cut&paste process causes errors because the file names are equal. If different file names would be used, the correct order of the pages will be lost. If the files would be renamed by using a suffix index, problems with gaps in the index sequence (caused by rescanned and deleted pages), will occur.

One possible way is to rename the files by using a prefix index number (Table 2). This can be done with a batch-renaming tool like *Total Commander* (www.ghisler.com).

Table 2 – Example of renaming odd files, the rescanned page n° 7 has no effect on the order of renamed files

Original page number in book	Old file name	New file name	Remark
1	~_0001.tif	0001~_0001.tif	
3	~_0002.tif	0003~_0002.tif	
5	~_0003.tif	0005~_0003.tif	
7	~_0005.tif	0007~_0005.tif	(Page scanned twice, ~_0004.tif moved to ~\scans\odd\bad_scans
9	~_0006.tif	0009~_0006.tif	
...

Figure 16 shows the renamed and merged pages ready for further workflow steps in WP5. Since all even pages are in an upside-down position, it is easy to find out if there are any errors by the renaming and merging process are caused.

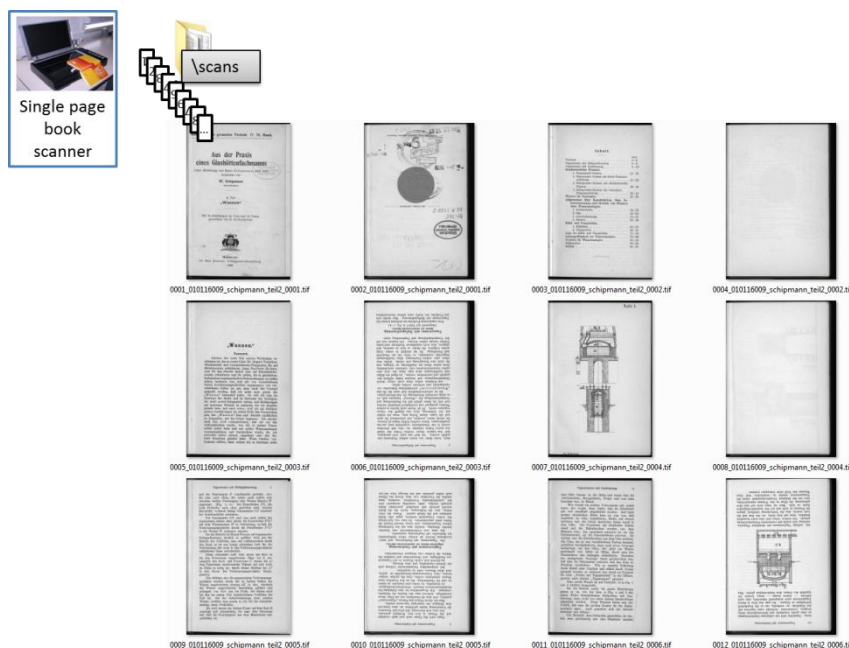


Figure 16 - All pages are renamed, merged and ready for processing

2.9 Quality Check Up

During the scanning process, a first quality check-up in the preview window of the scanner software for each page is necessary. It should be proved if the scanned page is complete, if the colour mode chosen correctly or if the page put straight on the scanner. If there is realized a faulty scan the page have to be rescanned again.

After finishing the scanning process, or merging process (if necessary), the quality, the correct order, the resolution, the colour mode, the non-compression file format and a plausible file size of the scans have to be checked randomly.

3 Results of the work package WP4 within the thinkMOTION project

In the first project year, regional digitisation service centres were established in all partner countries. It was necessary to perform training courses with the staff to practice the workflow, the handling with the database and to point to possible sources of errors and problems. In preparation to these training days a lot of training materials and a 50-page tutorial was written, which explain all necessary steps of the workflow for digitising of paper- based documents in detail. This tutorial summarises the experiences in a catalogue of rules, which can be used by other European digitisation projects (see Annex I).

The project partners acquired appropriate scanning devices. One partner has built a book scanner by itself another partner is designing its own book scanner now. All others own book scanners already, using it together with their libraries or bought ones. The thinkMOTION partners work together as a European digitisation network for this project and in the future for further European digitisation projects. By now more than 1,900 books, journal articles, teaching materials, over 3,600 slides, and more than 900 physical demonstration models could be digitized, processed in a quality improvement process and will be accessible via the Europeana online portal by the end of 2011.

While searching for relevant documents for the DMG-Lib portal some seldom documents were be found and digitised. Figure 17 shows two scanned pages of a handwritten transcript of the year 1868 of a lecture of Professor Franz Reuleaux, one of the most important scientists in mechanism science.

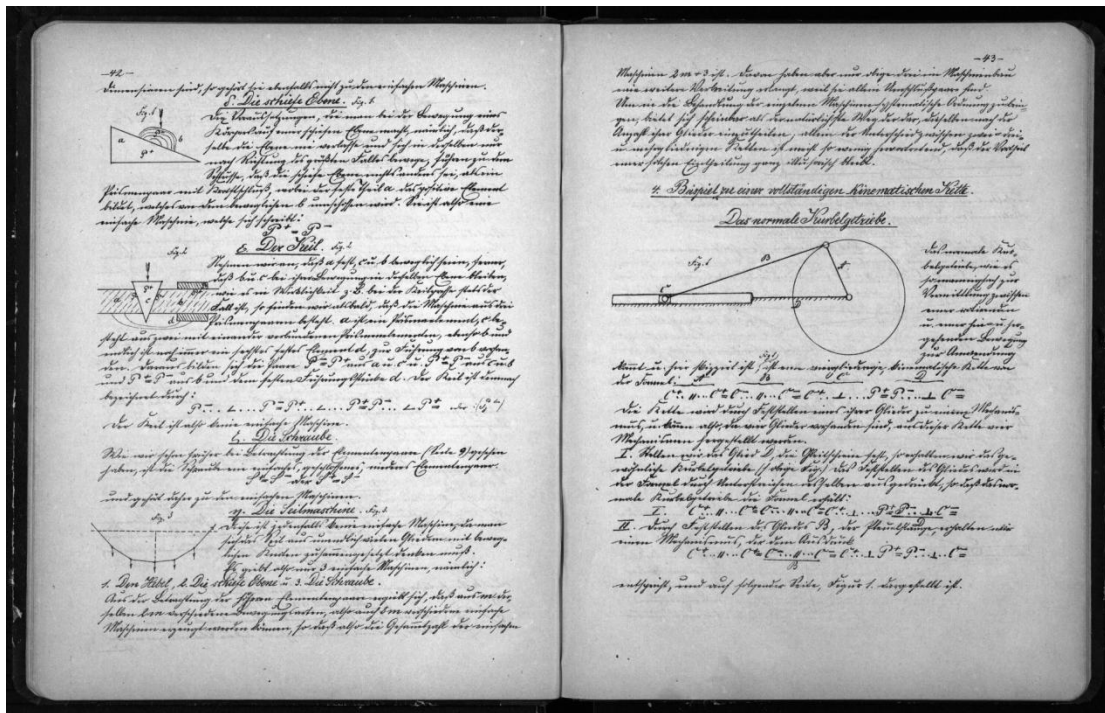


Figure 17. Unprocessed handwritten transcripts of 1868 of Franz Reuleaux's lecture "Vorträge über Maschinenbaukunde". Reuleaux is one of the most important scientists in mechanism science.

4 Outlook for the next project year

In the next project year, the number of digitized content will be increased. The scanning equipment basis and the workflow, especially for pictures, slides and physical demonstration models in motion, will be improved. Additionally training courses for these types of sources must be prepared and performed. Training materials and tutorials have to be worked out.

Annex I

Tutorial Workflow for Digitizing Paper Based Documents for the thinkMOTION Project



Tutorial

Workflow for Digitizing Paper Based Documents for the *thinkMOTION* Project

Authors: Veit Henkel, Erwin Lovasz, Michael Reeßing

Documents

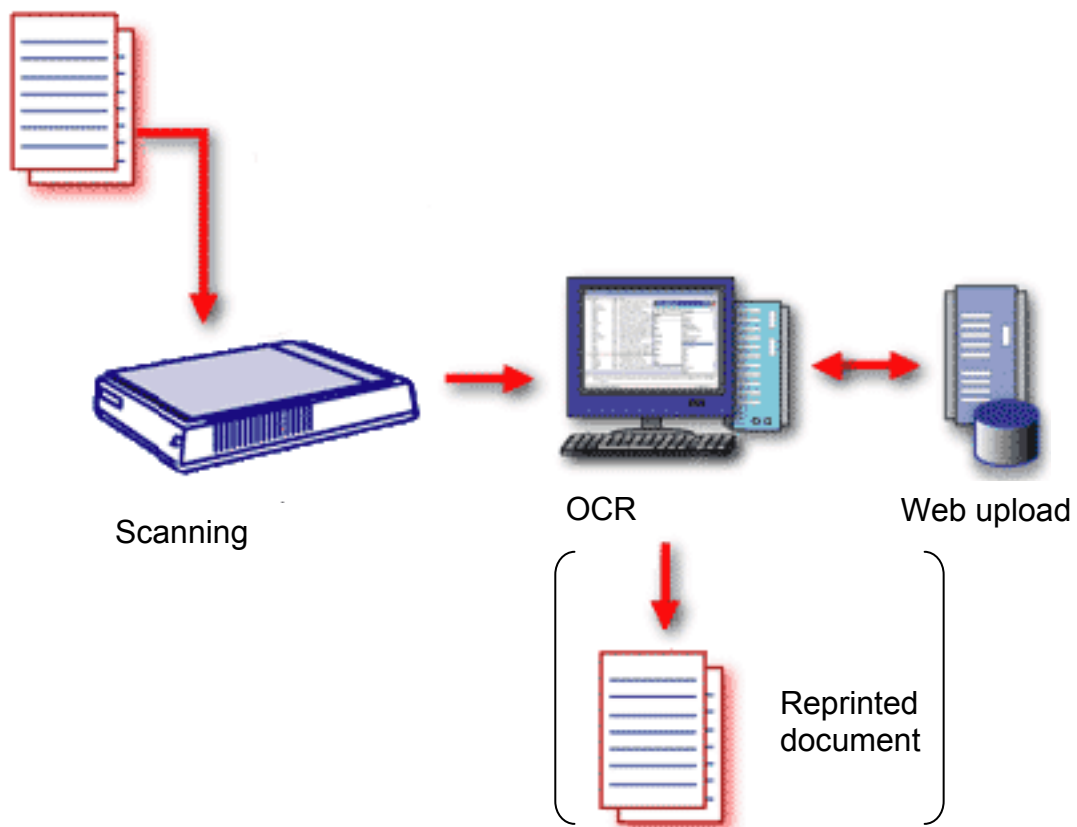


Table of Contents

1. Introduction.....	3
2. Data Handling – Part I: Reserve the Document on the ProDB Server and Create a Folder Structure on your Local Computer	4
2.1 Configuration.....	4
2.2 Description of the Scan Workflow Overview in the ProDB.....	5
2.3 Reserving a Document for Scanning and Processing	6
2.4 Description of the Folder Structure Created on your Local Computer	8
3. Scanning	9
3.1 Scanner Hardware	10
3.2 Scanner Software	11
3.3 Scanning Parameters	11
3.4 Saving your Scanned Raw Images.....	12
3.4.1 File Format.....	12
3.4.2 File Naming - Saving Path	12
3.4.3 Quality Check Up.....	17
4. Preparing and Quality Improvement of Scanned Documents	17
4.1 Overview	17
4.2 Download and Installation.....	18
4.3 Processing Steps for Preparing and Cleaning of Scanned Documents.....	18
4.3.1 Starting and Setting Up ScanTailor	18
4.3.2 Processing the Scanned Documents Step by Step in <i>ScanTailor</i>	19
5. Optical Character Recognition (OCR) of Scanned, Prepared and Cleaned Documents.....	32
5.1 Overview	32
5.2 Download and Installation.....	32
5.3 Processing Steps for OCR.....	32
6. Data Handling – Part II: Upload the Folder Structure of a Processed Document	35
Annex	38
Annex A – Check parameters for scanned files	38
Annex B – Examples for working with <i>ScanTailor</i>	40
Annex C – Examples for working with <i>ABBYY FineReader</i>	49

1. Introduction

Please note: The important keywords of the step descriptions in this tutorial are in **bold** letters, the names of buttons or fields, which are visible in the program windows (or screenshots) of the English software version as well as the names of the software, are in *italics*.

Figure 1 shows the workflow necessary to scan, process, and upload a document. This workflow involves two computers: the ProDB server and your local computer that is usually connected to the scanner hardware.

The four main steps are:

1. **Reserve the document on the ProDB server and create a folder structure on the local computer:** This will lock the document and shows other users of the ProDB that they should not scan it (to prevent redundant work). The folder structure will be created on the local computer and will hold the files for one document (scanned and processed content).
2. **Scanning:** Depending on the type of your scanner hardware you have to scan your files into a specific folder of the created folder structure by using a predefined file name on your local computer.
3. **Processing:** To improve the visual quality of your scanned images and to make the text of the documents searchable, some processing steps are necessary. The recommended software for processing is listed in Table 1.
4. **Upload:** As soon as scanning and processing is done on the local computer and the folder structure contains all necessary files, it can be uploaded to the ProDB server. There it will be archived, imported and linked to the corresponding database entries.

Table 1 - Recommended software for processing scanned documents

Name	Web address	Price	Remark
<i>IrfanView</i>	www.irfanview.com	Free, donations are welcome	Other picture viewer is possible.
<i>ScanTailor</i>	scantailor.sourceforge.net	Free, donations are welcome	Necessary
<i>ABBY FineReader</i>	www.abbyy.de	< 100 € in various online shops	Necessary
<i>ABBY FineReader XIX for Fraktur</i>	www.frakturschrift.com	0.04 ... 0.1 € per page	Only necessary if you have a lot of documents with Gothic type.
<i>Total Commander</i>	www.ghisler.com	Shareware, 28€	Other tool for batch renaming is possible.
<i>Adobe Professional</i>	www.adobe.com	About 100 € educational license	Helpful for quality checkup of your generated PDF files - but not absolutely necessary.

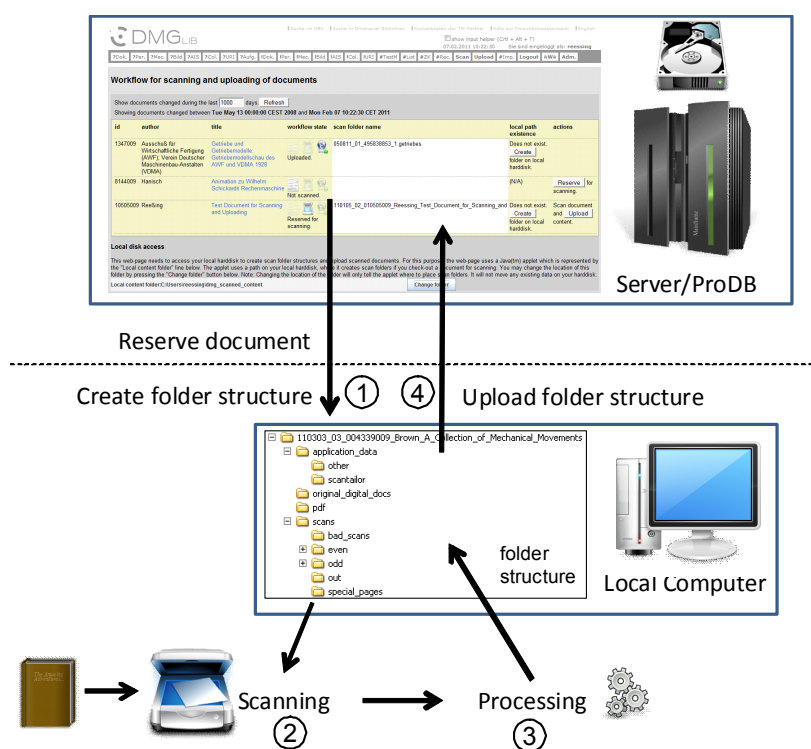


Figure 1 - General workflow for reserving and uploading documents

2. Data Handling – Part I: Reserve the Document on the ProDB Server and Create a Folder Structure on your Local Computer

2.1 Configuration

In this section you will learn how

- to grant permissions to the ProDB (*thinkMOTION Production Data Base*) to access your local hard disk,
- to specify the folder that will contain scanned content on your local hard disk.

The configuration steps have to be done only for the first time unless you start working with another computer. In such a case you have to repeat these steps.

Do the following configuration steps for the first time only:

- Navigate to the *Scan Workflow Overview* by pressing the **Scan tab button** in the ProDB main menu (Figure 3, Label 1; ProDB: <http://www.dmg-lib.org/dmglib/prodb/login.jsp>).
- The web browser will ask you if you allow a part of the ProDB website to receive extended permissions. These are necessary for the ProDB to create pre-defined folder structures on the hard disk of your local computer. Select **Run** (Figure 2). Note: You may also check the option *Always trust content from this publisher*. In this case the web browser will not ask for permissions each time you access the *Scan* tab.
- Specify a folder** that will contain scanned content on your local hard disk. For this, press the **Change folder** button (Figure 3, Label 2) and select or create a folder using the standard file open dialog. A good choice would be "c:\scan". The ProDB website will create folders for each scanned document into this base folder. (Please note: You have to specify this path again if you change your PC.)

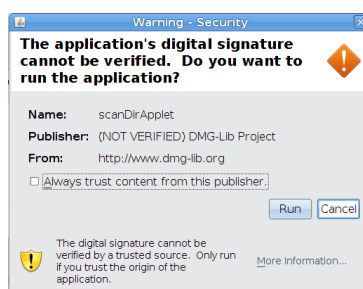


Figure 2 - Grant permissions to ProDB website

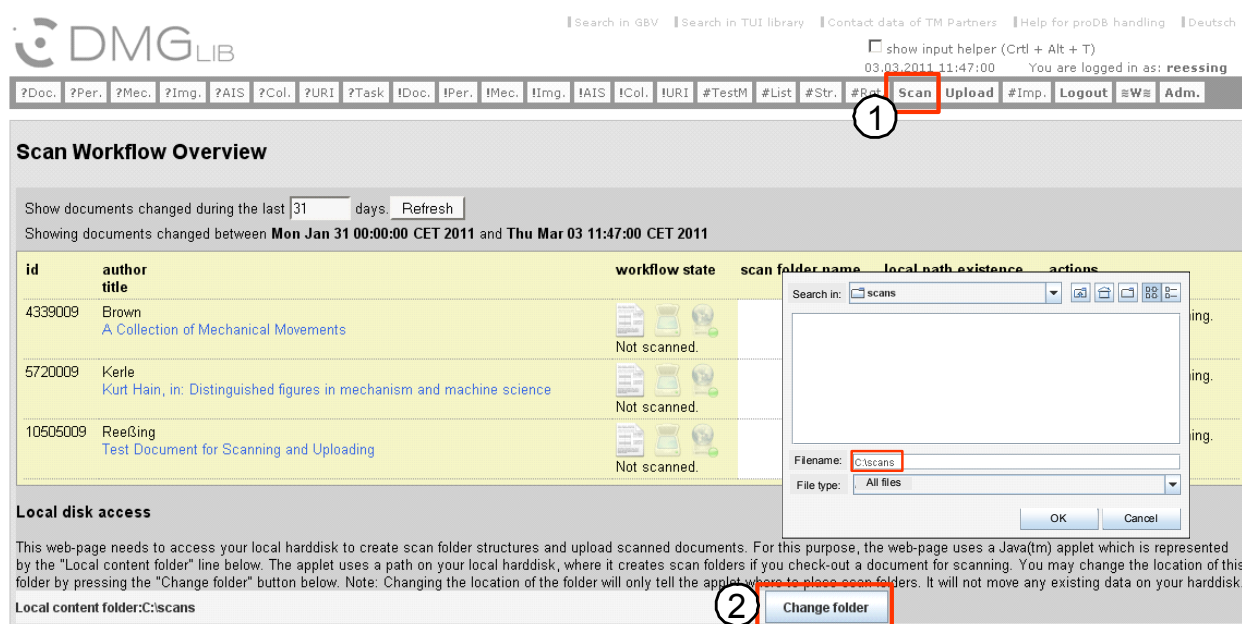


Figure 3 - Configuration of Scan Workflow Overview

2.2 Description of the Scan Workflow Overview in the ProDB

The Scan Workflow Overview displays a table of documents that you have edited during a number of recent days (Figure 4, Label 1). Each line of the table contains the **DMGLib ID**, **author and title** and the **workflow state** of one document (Figure 4, Label 2).

The state can be

- **Not scanned.** – The document has not been scanned or reserved for scanning.
- **Reserved for scanning.** – The document has been reserved and will be scanned and processed soon.
- **Uploaded.** – The document has been scanned, processed and uploaded.

Other columns of the workflow table are the **scan folder name** the server has assigned to this document and a column (**local path existence**) that indicates if this scan folder exists on your local hard drive in the content folder (see Configuration section above). If a scan folder name has been assigned but does not exist locally, the column contains a button to **Create folder on local hard disk**.

The last column lists available **actions** that depend on the state of each document. These actions are

- **Reserve for scanning.** – By pressing this button you reserve the document for scanning and processing. The document is then locked and will stay so until you upload the scanned and processed content.

- **Upload content.** – As soon as you have put the raw data and processed content into the folder structure, you may press this button to upload the whole folder structure to the server.

DMG LIB

Search in GBV Search in TUI library Contact data of TM Partners Help for proDB handling Deutsch

show input helper (Ctrl + Alt + T)
03.03.2011 11:47:00 You are logged in as: reessing

?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI ?Task ?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI #TestM #List #Str. #Rgt. Scan Upload #Imp. Logout ≡W≡ Adm.

Scan Workflow Overview

Show documents changed during the last 31 days Refresh **1**

Showing documents changed between **Mon Jan 31 00:00:00 CET 2011** and **Thu Mar 03 11:47:00 CET 2011**

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	Not scanned.		(N/A)	Reserve for scanning. 2
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	Not scanned.		(N/A)	Reserve for scanning.
10505009	Reeßing Test Document for Scanning and Uploading	Not scanned.		(N/A)	Reserve for scanning.

Local disk access

This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk.

Local content folder:C:\scans [Change folder](#)


Figure 4 - Workflow state and reserving documents

2.3 Reserving a Document for Scanning and Processing

Before you scan a document, you need to reserve this document on the server so that no one else tries to scan it at the same time. Follow these steps to reserve a document.

Please follow the instruction:

- Navigate to the **Scan Workflow Overview**
- Find the table entry for the document you would like to scan. Check that its **workflow state** is **Not scanned** (Figure 5, Label 1) and that the **Local Content folder** is set correctly.
- **Press** the **Reserve** button in the **actions** column for the document (Figure 5, Label 2). A pop-up will ask you if you are sure that you want to reserve the document since it requires you to scan, to process and to upload the content of the document (Figure 6, left). Press **OK** to continue.
- Another pop-up will inform you if the reserve request was carried out successfully (Figure 6, right).






☐ show input helper (Ctrl + Alt + T)
03.03.2011 11:47:00 You are logged in as: reessing

?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI ?Task IDoc. !Per. !Mec. !Img. !AIS !Col. !URI #TestM #List #Str. #Rgt. Scan Upload #Imp. Logout ≡W≡ Adm.

Scan Workflow Overview

Show documents changed during the last days. [Refresh](#)

Showing documents changed between **Mon Jan 31 00:00:00 CET 2011** and **Thu Mar 03 11:47:00 CET 2011**

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	<div>1</div>  Not scanned.		(N/A)	<div>2</div> <input type="button" value="Reserve"/> for scanning.
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	 Not scanned.		(N/A)	<input type="button" value="Reserve"/> for scanning.
10505009	ReeSing Test Document for Scanning and Uploading	 Not scanned.		(N/A)	<input type="button" value="Reserve"/> for scanning.

Local disk access


This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk

Local content folder: C:\scans

Figure 5 - Buttons for reserving a document that is in the state *Not Scanned*.

Figure 6 - Question if the user is sure that he wants to reserve the document (left) and result of the reserve request (right)

The result of the reservation process is shown in figure 7. The column **workflow state** has now changed to **Reserved for scanning** and the column **scan folder name** contains the folder name that has been determined by the server. Additionally, the column **local path existence** indicates if this folder exists in the base folder on your local hard disk (e.g. “c:\scans”, see section Configuration). After pressing the **Reserve button**, the folder structure should be created automatically on your local PC.



Search in GBV

Search in TUI library

Contact data of TM Partners

Help for proDB handling

Deutsch

☐ show input helper (Ctrl + Alt + T)

03.03.2011 11:47:00

You are logged in as: reessing

?Doc.

?Per.

?Mec.

?Img.

?AIS

?Col.

?URI

?Task

!Doc.

!Per.

!Mec.

!Img.

!AIS

!Col.

!URI

#TestM

#List

#Str.

#Rgt.

Scan

Upload

#Imp.

Logout




≡W≡

Adm.

Scan Workflow Overview

Show documents changed during the last days. [Refresh](#)

Showing documents changed between **Mon Jan 31 00:00:00 CET 2011** and **Thu Mar 03 11:47:00 CET 2011**

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	 Reserved for scanning.	110303_03_004339009_Brown_A_Collection_of_Mechanical_Movements	Exists on local harddisk.	Scan document and Upload content.
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	 Not scanned.		(N/A)	Reserve for scanning.
10505009	Reeßing Test Document for Scanning and Uploading	 Not scanned.		(N/A)	Reserve for scanning.

Local disk access

This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk!

Local content folder: C:\scans

[Change folder](#)

Figure 7 - Successfully reserved document

2.4 Description of the Folder Structure Created on your Local Computer

By means of pressing the *Reserve* Button in the ProDB a folder structure (Figure 8) will be created on your local computer in the specified *Local content folder* (Base Folder).

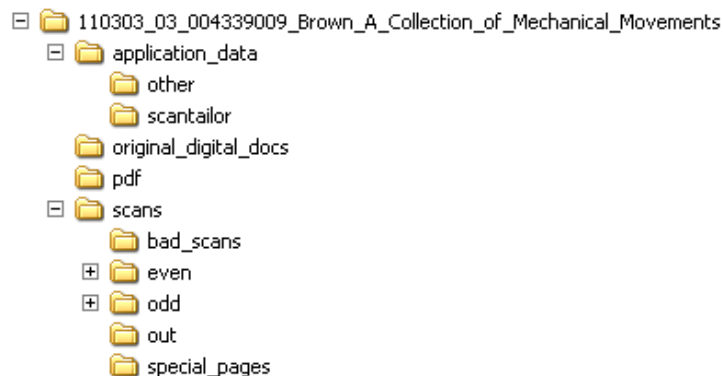


Figure 8 - Example for the predefined folder structure created on your local hard disk – Do not change these folder names!

The following name scheme is for your information only, you are not allowed to change the names and folder structure generated by the database. The names of the folders means:

Main folder name: yymmdd_##_xxxxxxx_authorsname_begin_of_title (generated by ProDB)

yymmdd	Date of starting process (2 digits for year, month and day)
##	Serial number of source of the day (start index = 01 for each day)
xxxxxxx	9-digit DMG-ID (creating by ProDB)
authorsname	Name of the main author or editor
begin_of_title	The first significant words of the title

Subfolders used for:

("~/") means the path you have used for download the folder structure from the ProDB)

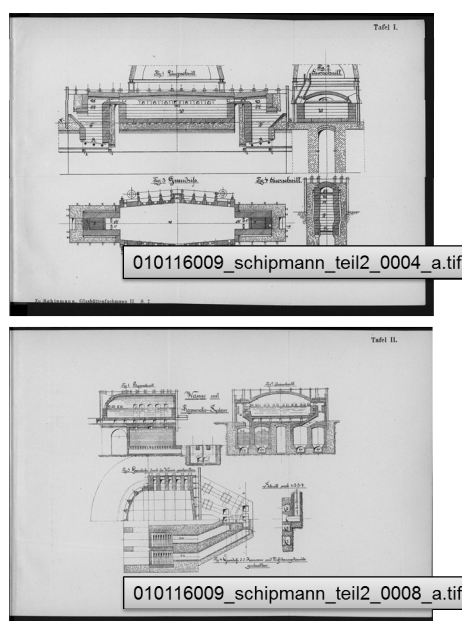
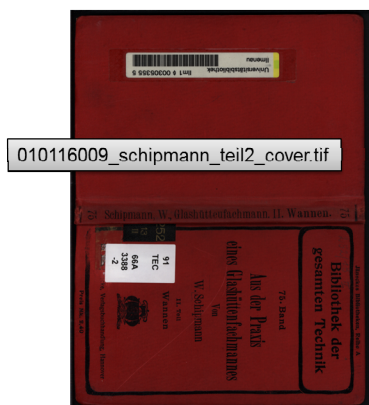
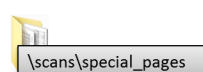


Figure 9 – Special pages such as a cover page and two fold-out pages scanned in the fold-out state

Table 2 – Meaning and intention of the folders and subfolders generated by the ProDB

Subfolder name	Description what folders contain
~\application data	Project files of tools used such as <i>ScanTailor</i>
~\pdf	Output of the OCR tool <i>ABBYY FineReader</i> as PDF/A-File Do not use this folder for other PDF documents such as described in ~\original_digital_docs!
~\original_digital_docs	Original files, which already exist as files and do not need a scanning process such as MS Word or PDF documents
~\scans	Scanned and unchanged images, the so called raw scans, as uncompressed or lossless compressed files. Use this folder for scans from a double page book scanner, from flatbed scanner or from a sheet feeder scanner. Use it also for fold-out pages scanned in the fold-in state.
~\scans\bad_scans	Scanned images, which are incorrect and rescanned again – please move this images into this folder and do not delete it.
~\scans\even	See ~\scans But use this folder for scans of the even pages (2,4,6,8, ...) if you use a single page book scanner (Figure 15).
~\scans\even\bad_scans	See ~\scans\bad_scans but for scans of even pages
~\scans\odd	See ~\scans But use this folder for scans of the odd pages (1,3,5,7, ...) if you use a single page book scanner (Figure 15).
~\scans\odd\bad_scans	See ~\scans\bad_scans but for scans of odd pages
~\scans\out	Processed output images of the quality improvement tool (<i>ScanTailor</i>)
~\scans\special_pages	Scanned images of special pages such as fold-out pages scanned in the fold-out state, cover pages if in the original state, etc. (Figure 9) Images in this folder are the unchanged raw scans as uncompressed or lossless compressed files.

3. Scanning

The aim of this workflow step is to make image files from paper sources (books, articles, etc.) and to store them in a given folder structure according to a given name scheme.

Because of further processing steps all pages have to be scanned, also **blank pages!** **Cover pages** should be scanned if they are in the original state and not blank (Figure 9).

“**Fold-out pages**” have to be scanned within the fold-in state in the normal scan sequence. After finishing the process by scanning the last page, the fold-out pages have to be scanned in the fold-out state again (Figure 9). If those are too large for one scan, then scan them in multiple overlapping sections (naming see below “File naming”).

Scan only documents for which you have the chance to clarify the rights of use!

3.1 Scanner Hardware

Generally there are three different types of scanners concerning the types of output files:

- Double page book scanner (delivers a sequence of images containing two pages in each file)
- Single page book scanner (delivers a sequence of images containing single pages, only one type of pages, the even or the odd pages are in an upright position)
- Flatbed and sheet feeder scanner (delivers a sequence of single pages in an upright position)



Advantages:

- scanning area up to DIN A1 (594mm x 841mm)
- gentle to books
- easy to handle also for heavy books
- double page scanning mode

Disadvantages:

- rather expensive
- big (about 4m² area), heavy



Advantages:

- cheap
- small, light

Disadvantages:

- hard work if heavy books (> 200 pages) have to be scanned
- not suitable for books with broken bindings
- single page scanning mode



Advantages:

- very fast
- double side scanning mode
- works automatically

Disadvantages:

- not suitable for bound documents

Figure 10 – Double page book scanner *Zeutschel* with a special book cradle (left); single page book scanner *Plustec A300* (middle); sheet feeder scanner *HP Scanjet* (right)









Type of scanner	Scanner (software) output
 <p>Single page book scanner</p>	<p>Odd and even pages in separate directories; even pages are upside-down</p> <p>(Alternated scanning of odd and even pages in one directory: Book has to be turned permanently. It is a hard work and a risk of double scanned pages or forgotten pages.)</p> <div>   </div>
 <p>Double page scanner</p>	<p>Two pages in each file</p> <div>  </div>
 <p>Sheet feeder scanner</p>	<p>One page in each file</p> <div>   </div>

Figure 11 – Output of the scanner software depending on the type of used scanner hardware

3.2 Scanner Software

Usually each scanner will be delivered with its own scanner software. In this tutorial it is not possible to describe each software tool individually.

In some cases the delivered scanner software does not work properly. E.g. the book scanner *Plustek OpticBook A300*, by itself a good device, the software has a bug. In the *Scan-to-file-mode* different picture formats can be chosen, such as *TIFF-uncompressed* without any parameters and *TIFF-compressed* with compression parameters – but both settings save your scan files in a JPEG-compressed TIFF-file with a loss of quality. *ScanTailor*, the tool for the next step in the workflow, cannot handle this file format. It shows only black images after the importing process. Alternatively, you can use a picture viewer with a batch scanning function, such as *IrfanView*. In this case you have to install at least the vendor's scanner driver. You can start the batch scanning process by choosing your scanner in the *File/Select TWAIN source* (not the WAI device) then chose *File/AcquireBatch-Scanning*.

If you use the *Plustek OpticBook A300* Scanner via TWAIN interface with *IrfanView*, you cannot use the buttons at your scanner device and you have to trigger your scanning process with the mouse button or the keyboard. It is not an easy handling. The problem is: You have to press the book against the scanner's book edge, normally with two hands, in order to get a good result and you have to start scanning by mouse or keyboard at the same time. A well-proven solution is a foot switch in addition to mouse and keyboard. So you have both hands on the book and you can trigger the scanning process with your foot. It is even easier than the device buttons.



Figure 12 – Foot switch for triggering the scan process

3.3 Scanning Parameters

Since each scanner software has its individual user interface and parameters it is not possible to write down a tutorial for this. From the experience with different scanner software some hints will be given below. It is your responsibility to find the positions in your software where you can set the parameters.

The following parameters have to be observed:

- **Colour Mode:** Scan colour mode depends on your type of document
 - Colour mode with 24 bit, “True Colour”, 16,7 Million colours for documents with more than about 5% coloured pages
 - Grey scale with 8 bit, 256 grey scales, for all black and white or grey shaded documents
 - Scan all pages with grey scale and additionally all colour pages in colour mode if the document contains about less than 5% colour pages
 - Never use 2 bit, bitmap or black and white scan mode! Even though in later steps we will convert it to black and white the scanned raw images must be archived in colour or grey scale mode (preservation).
- **Resolution:** Scan generally with a resolution of **at least 300 dpi**. Documents with small lines and small fonts should be scanned with 400 dpi. Do not change the resolution within a document! It causes problems in later steps.

- **Do not use any image quality improvement tools:** In normal case do not use any kind of filters and image corrections functions such as brightness, contrast, sharpness etc.
- **Scanning area:** You can reduce the file size if you reduce the scanning area. But make sure that no content disappears. Use the same area (width and height) for the same size of pages within a document. Reducing the scanning area can also speed up the scanning process significantly.

File size of the scanned images can reach approximately between 5 and 70 MB depending on the resolution, the colour mode and the page size. And so you need adequate space at your hard disk.

3.4 Saving your Scanned Raw Images

3.4.1 File Format

Scan files must be saved in a **TIFF uncompressed** file format (lossless compression as TIFF-LZW or TIFF-ZIP is possible to reduce the file size but not recommended). Never use JPEG compression!

3.4.2 File Naming - Saving Path

Table 2 gives you an overview of the meaning of the folders automatically generated by the ProDB website. The next two chapters give you more detailed information.

Please make sure you use the following file name scheme for naming your scanned files. Normally your scanner software automatically generates file names suffixes (running numbers). Please note the start index, the number o digits and the increment value!

File naming: xxxxxxxxx_####.tif

xxxxxxx	9-digit DMG-ID (see DMG-Lib – ProDB, the same like in folder name)
####	4-digit serial number with the following parameters:

- start index = 0001;
- increment for the next single or double page image = 1;
- index for scanning of **fold-out pages** in the fold-out state = “page index of the fold-out page scanned in the fold-in state”+[a-z],

Example:

- image of the fold-out page scanned in the fold-in state:
001547009_0067.tif,
- image of the fold-out page scanned in the fold-out state
001547009_0067a.tif (Figure 9), and if it is too large for your scanner hardware, scan it into multiple images and attach b, c, d ... to the file name such as 001547009_0067b.tif ...

Depending on your type of scanner hardware (Figure 10) and the type of scanned pages different folders (Figure 8) should be used for saving scanned images. The scanned images have to be saved into the folder *scans* or in the corresponding subfolder generated by the “Reserving a Document for Scanning and Processing”-process described above. Figure 14 shows the general workflow for processing documents and especially the different handling for the output of different scanners.

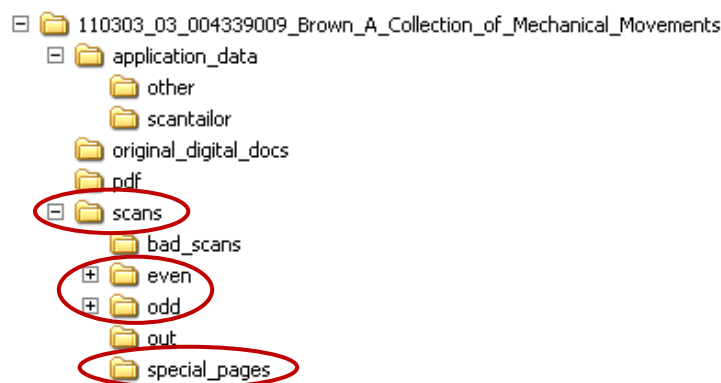


Figure 13 – Folders which should be used for saving scanned images, depending on the type of scanner and the types of pages

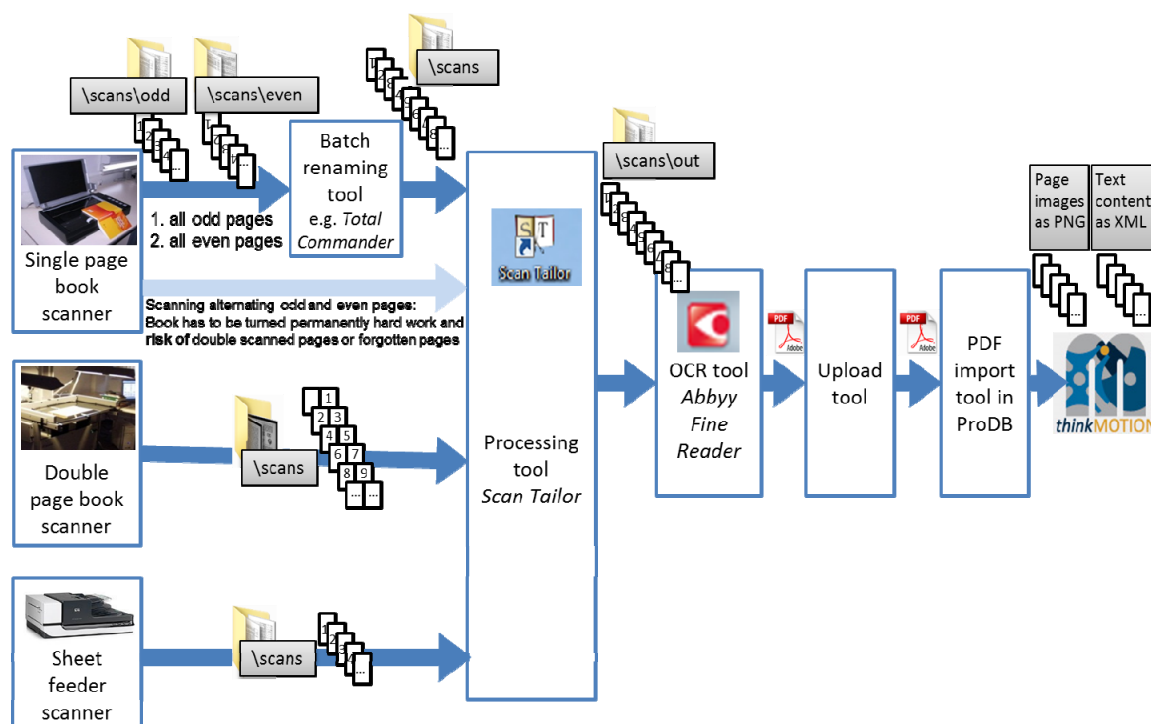


Figure 14 – General workflow for processing documents, folders for saving scanned images depending on the type of scanner hardware and the flow of generated documents

If you use a sheet feeder or a flatbed scanner or double page book scanner, you should save your scanned images directly to the “**scans**” folder. If you use a single page book scanner, scan in a first run all odd pages into the folder `~\scans\odd`. In a second run scan all even pages to the folder `~\scans\even` (Figure 15). See also chapter 3.4.2.2 *Additional File Handling for „Single Page Book Scanner“* for further explanations and necessary additional steps.

If your scanned image is not acceptable and you would like to rescan a page again, **please follow the instruction:**

- If you have **realized** the fault immediately **after scanning this page** and before scanning the next page, **move** the **incorrect page** to the subfolder **bad_scans** and **rescan** the **page** again. Please delete the incorrect pages only if you rescan the whole book. It does not matter if the file index has a gap. It is not necessary to change the start index in your software. Important is only that the file index is increasing corresponding to the original order of your document pages.
- If you have **realized** the fault **after scanning the next pages** or the last page you have to **move** the **incorrect page** to the subfolder **bad_scans** and **rescan** the **page** again. Then you have to **rename** the file to the file name of the incorrect and moved page.

3.4.2.1 Additional File Handling for Special Pages like Cover or Fold-Out Pages

Cover pages, if they are not blank, and fold-out pages in the fold-out state should be scanned into the folder ~\special_pages. These pages have to be processed separately from the normal content pages. After finishing the quality improvement workflow step (with *ScanTailor* tool) and before starting the OCR process with *ABBYY FineReader*, you have to merge these special pages with the normal scanning sequence.

The fold-out pages in the fold-out state should be inserted after the fold-out pages in the fold-in state scanned in the normal scanning sequence. So the user can see, while page turning in the portal, these pages first in the fold-in state and then in the fold-out state, just as in the original document. Before inserting these pages, please observe the correct file naming as defined above, so that the pages are in the correct alphabetic order as mentioned previously. Check in any case the correct order after inserting all special pages before you continue.

3.4.2.2 Additional File Handling for „Single Page Book Scanner“

For single page book scanners, such as the Plustek OpticBook A300, scanning of odd and even pages alternately (means scanning following the page numbers) leads to errors such as twice scanned pages or forgotten and not scanned pages, because the book has to be turned after each scan. Another problem is the additional physical work load for the operator by turning the book permanently.

For this reason it is recommended to scan all odd pages first and then all even pages in different folders.

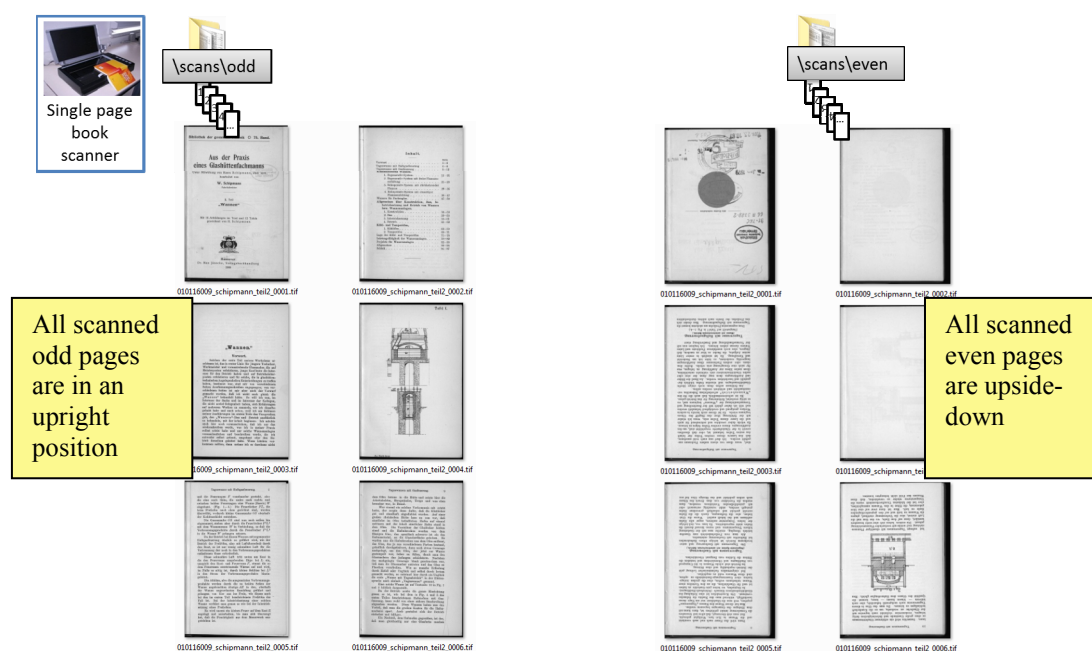


Figure 15 – Example of scanned pages in the odd and the even folder

The scanned odd pages are in an upright position and all even pages are upside-down and have to be turned in a further workflow step with the tool *ScanTailor* (Figure 15).

After finishing the scanning process with a single page book scanner you have to merge both scanning sequences to one in the folder `~\scans`. A simple cut&paste process causes errors because the file names are equal. If we would use different file names we lost the correct order of the pages. If we rename the files using a suffix index, we will get problems with gaps in the index sequence as a result of rescanned pages.

One possible way is to rename the files by using a prefix index number (Table 3). This can be done with a batch renaming tool like *Total Commander*.

Table 3 – Example of renaming odd files, rescanning page n° 7 has no effect on the order of renamed files

Original page number in book	Old file name	New file name	Remark
1	~_0001.tif	0001~_0001.tif	
3	~_0002.tif	0003~_0002.tif	
5	~_0003.tif	0005~_0003.tif	
7	~_0005.tif	0007~_0005.tif	(Page scanned twice, ~_0004.tif moved to ~\scans\odd\bad_scans
9	~_0006.tif	0009~_0006.tif	
...

For merging odd and even pages please follow the instruction:

- Download and install the software tool *Total Commander* from www.ghisler.com
- **Open the *odd* folder**
- **Select all images** (Select first image, hold down SHIFT key, select last image – all images have to be red coloured)

- Chose **Files/Multi-Rename-Tool ...**

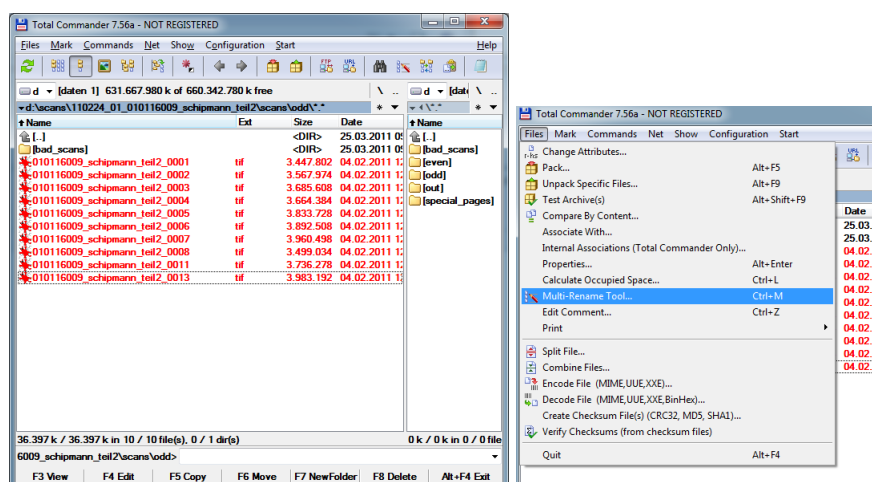


Figure 16 – Total Commander – selecting all images and starting renaming tool

- Change the **settings** like this (see also figure 17):
 - *Rename mask:* [C]_[N] (C: counter; N: old file name)
 - *Define counter [C]:*
 - *Start at:* 1
 - *Step by:* 2
 - *Digits:* 4
- Check the **result** at the preview
- Save **settings** as: “**odd**” (for later use – you can load it with F2-key)
- Press **Start!** button

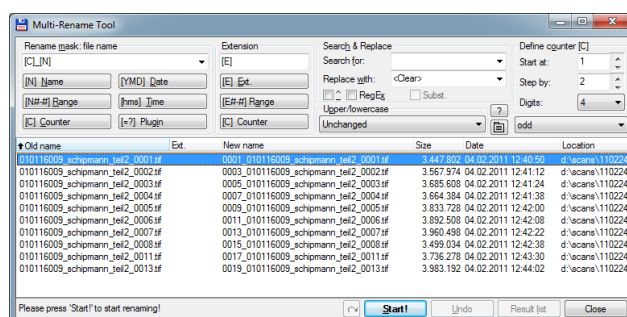


Figure 17 – Total Commander - settings in the batch renaming tool

- Do the **same with the even** folder with the these changes:
 - *Define counter [C]:* *Start at:* 2
 - **Save settings** as: “**even**” (for later use)
- **Copy all files** from the *odd* and the *even* folder **to** the upper **~lscan** folder
- **Check the correct order** randomly with a picture viewer like *IrfanView*, but at least the first five pages, five pages from the middle and five pages at the end of the document. The page numbers have to be in the correct order at any position in the document (Figure 18). Otherwise you have to find out the reason. (Often not scanned blank pages causes problems on merging odd and even pages)

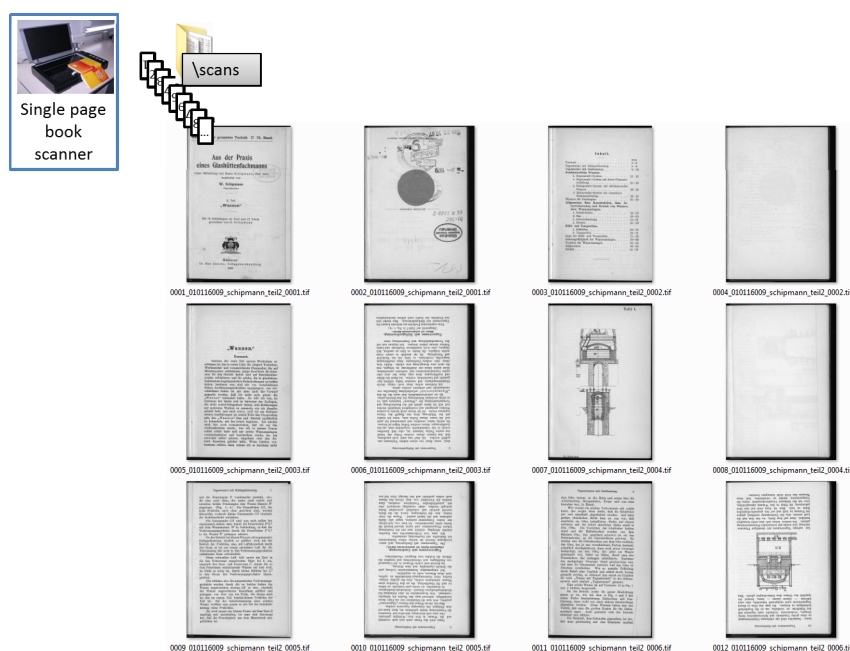


Figure 18 - All pages are renamed, merged and ready for processing

3.4.3 Quality Check Up

During the scanning process you should check each page in the preview window of your scanner software or your picture viewer like this:

- Is the scanned page complete?
- Is the colour mode chosen correctly?
- Is the page laid straight on the scanner?

If you realize a faulty scan rescan the page immediately as described above.

After finishing the scanning process, or merging process (if necessary), check the quality of your scans, the correct order, the resolution, the colour mode, the non-compression and a plausible file size randomly.

4. Preparing and Quality Improvement of Scanned Documents

4.1 Overview

Please note: The name of the files and the screenshots in this tutorial do not correspond in all cases. The screenshots show partially different content sources in order to illustrate all important cases and facts.

The aim of the processing steps, described in this part of the tutorial, is to improve the visual quality of the scanned images with a reasonable effort. In the *thinkMOTION* project we recommend to use for this task a free software tool called *ScanTailor*. *ScanTailor* is an interactive post-processing tool for scanned pages and performs operations such as page splitting, deskewing, adding/removing borders, and others.

The inputs for *ScanTailor* are the scanned raw image files. *ScanTailor* delivers processed output files as LZW lossless compressed TIFF. These processed files are the input for the OCR process, which is described in the next part of the tutorial.

ScanTailor works nearly automatically, but you have always to check up the results and you have often to change setups manually with more or less effort. The following aims should be achieved in this process before starting the OCR processing step and the following hints should be observed:

- The pages should be in the correct upright position like the original
- Double pages have to be split into two single pages
- Blank pages must not be deleted
- The original order of the pages must not be changed
- The size of the margins should be defined in this way that the page layout looks good.
- The margins should be free of parts of background not belonging to the page (e.g. book cradle or cap from the scanner)
- The text blocks should be aligned horizontally respectively vertically and should not jump on turning pages
- The types (letters, numbers etc.) should be well readable
- No part of the content must disappear.
- Lines and other fine structures in drawings should be visible again
- The content of photographs should be well recognisable
- For colour scans: the colour should be close to the original

4.2 Download and Installation

First of all you have to download and install the *ScanTailor* software.


Please download the latest version from:

<http://scantailor.sourceforge.net>

4.3 Processing Steps for Preparing and Cleaning of Scanned Documents

4.3.1 Starting and Setting Up ScanTailor



After installing **start *ScanTailor*** using the icon  at your desktop or chose it from Start/Program button at your Windows task bar. The program window opens and you can choose between *New Project* and *Open Project*. New projects or access to existing ones is also possible from the menu *File*, submenu *New Project* or *Open Project*.

Please follow the instruction:

- Select **New project**
- Press **Browse** button to choose the **Input Directory**
- **Select path** where your raw scan images are saved and press **OK** button
(E.g.: C:\scans\110125_05_010607009_reuleaux_cultur_und_technik\scans)
- **Check** whether all TIFF files, which are to be processed, are in the field *Files In Project* and whether only scan process data files or special scan files are in the *Files Not In Project* field. Only the “files in project” are processed!
- Leave the **Output Directory** unchanged at ~/out and press **OK**

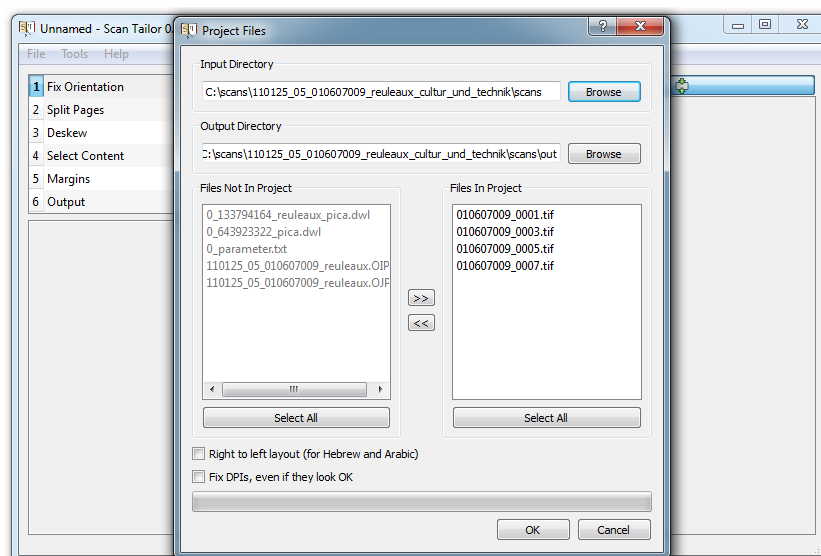


Figure 19 – ScanTailor project configuration window

4.3.2 Processing the Scanned Documents Step by Step in ScanTailor

In the upper left corner of the *ScanTailor* window is the navigation section to navigate through the preparing and cleaning steps. The steps, numbered from 1 to 6, are in detail:

1. *Fix Orientation*
2. *Split Pages*
3. *Deskew*
4. *Select Content*
5. *Margins*
6. *Output*

Usually you have to go step by step through these points from 1 to 6. But it is also possible to repeat each step for a single or for all pages without doing the following steps again. It can be necessary if you find an error in a later step. E.g. you see that the automatic page splitting did not work correctly, you can reposition the cutting line manually.

After importing scanned images into your project the first raw scanned image in the selected directory is displayed in the main window. On the right side of the *ScanTailor* window, a preview of all loaded scanned images is displayed as “thumbnails”. You can click at a thumbnail picture to jump to a specific image. It is also possible to remove images from the project by pressing the right mouse button if you have imported a single wrong image or a fold out page which should be processed in a new *ScanTailor* project.


Please note that the most of the settings at each step have only an effect to the current page. If the settings are the same for all pages, you have to apply it to all pages. But some settings have to be done for one page only. Please execute each step according to the following description.

Step 1 – Fix Orientation

If necessary, the page orientation can be modified with left or right rotation in 90° steps.

If you work with a single page scanner, you normally get a sequence of scanned pages where the odd pages are upside down and the even pages are in the upright position. In this case you have to turn all odd pages by 180 degrees.

Please follow the instruction:

- Choose the **first wrong oriented page**
- Press the **rotation buttons**: once for 90° and twice for 180°
- Press **Apply to...** and **select** in the window *Fix Orientation* the **correct option**:
 - Use Apply to **All pages** if all pages are turned wrong in the same manner.
 - Use Apply to **Every other page** to turn only the odd or only the even pages (For a correct function it is necessary that all even and all odd pages are scanned and imported into the project, also the blank pages!)
 - If single pages are wrong oriented, rotate them individually page by page.
- Press **OK**
- **Go to the first image** by clicking at the first thumbnail picture (Otherwise the batch process runs from the selected page to the last page only.)
- Push the **play** button  to launch the batch process for all scanned files.

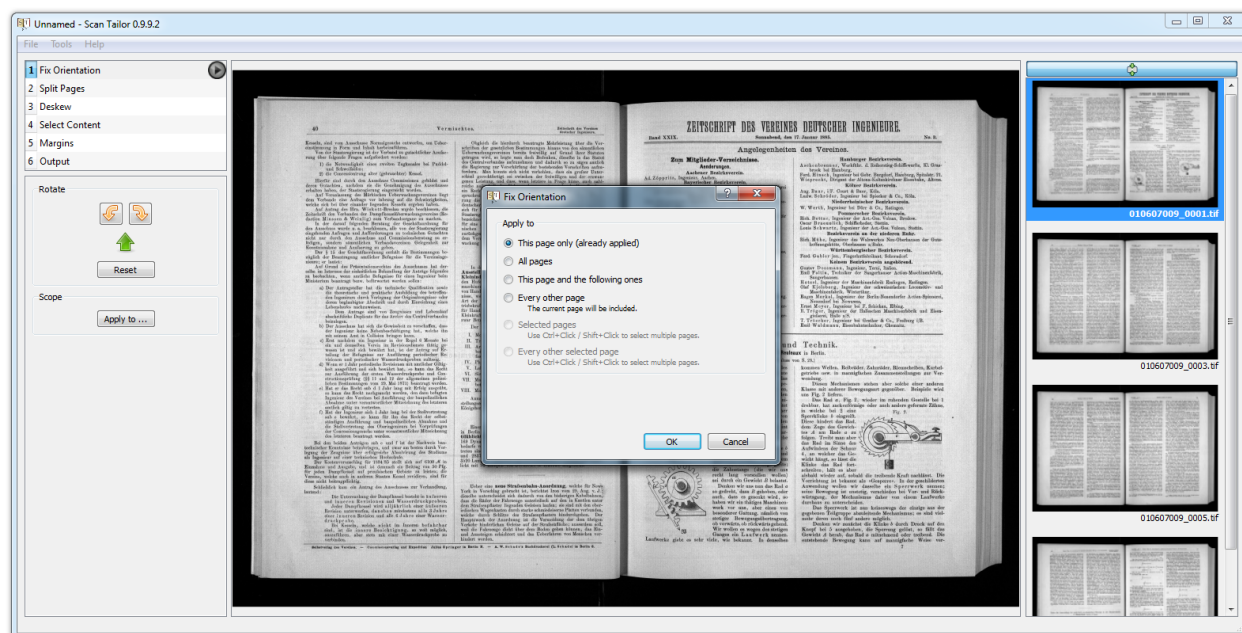


Figure 20 - ScanTailor with the settings for the orientation of the images

Step 2 – Split Pages

This step allows splitting or cutting your scanned pages. The splitting/cutting mode depends on the type or layout of your scanned pages and/or the type of your scanner hardware. In *ScanTailor* you can choose between the following three cases of page layouts:

- If you use a double page scanner, the scanned file contains the images of two pages and has to be cut between these two pages in the book edge.

- Another case is the following. You have scanned each page separately but there is the current page and a piece of the previous or the next page including the book edge in your image. So you can cut away the unrelated piece of the other page. Common flatbed scanners deliver these types of scans.
- If you use a flatbed scanner with a special book edge, you have normally only one page per file and so you do not need any cutting or splitting.

ScanTailor can find the splitting/cutting position automatically. You can change it manually if necessary. For this the cutting line(s) can be moved by pushing the left mouse button and dragging to the desired location, towards right or left (see figure 21).

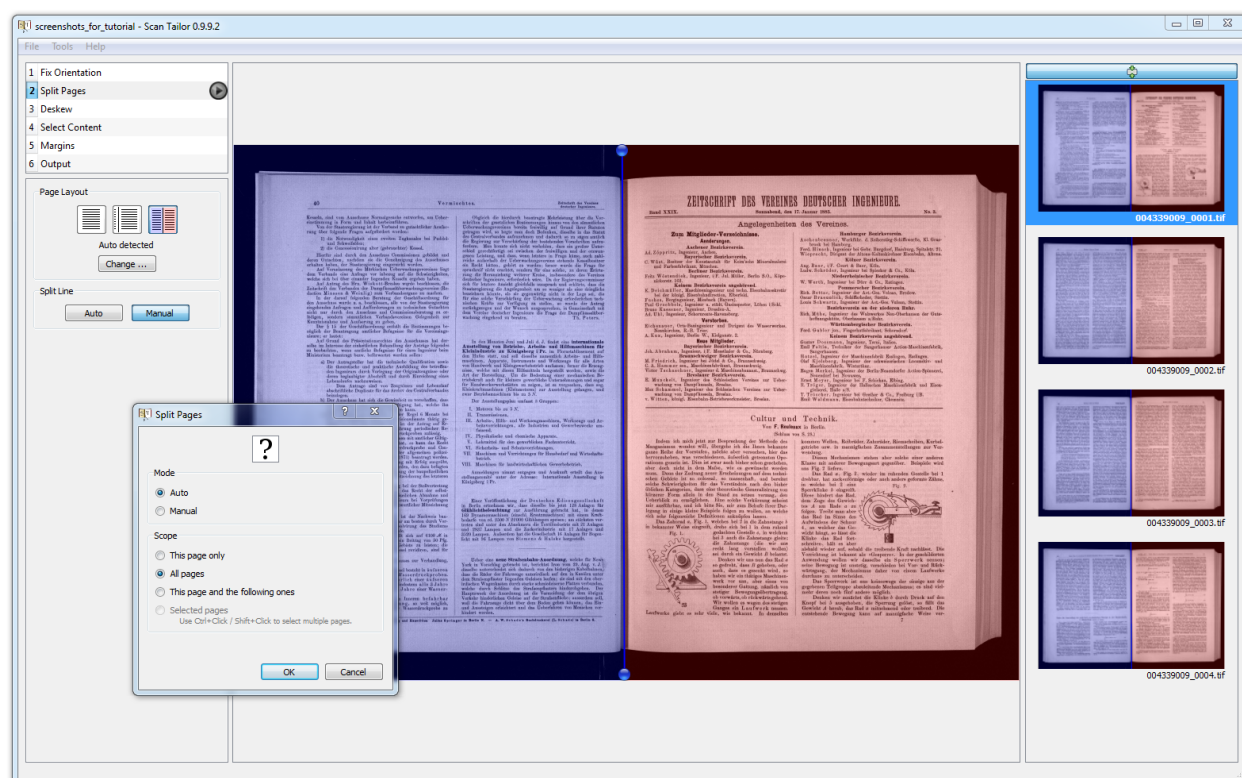



Figure 21 - *ScanTailor* - splitting a double page scanner image into two pages

Please follow the instruction:

- Choose the **type of your scanned images** by pressing the corresponded icon
- Press **Change** button
- Select the option **Auto** and set the scope to **All pages** to make it valid to all pages and press **OK**
- Go to the **first image** by clicking at the first thumbnail picture (In other case batch process runs from the selected page to the last page only.)
- Push the **play** button  to launch the batch process for all scanned files.
- Check **visually all pages** by browsing through all thumbnails to check up the result of automatic mode or use Page up/Page down keys
- If there are images where the pages are **not split correctly**, set the **position manually**

Step 3 – Deskew

In this step you can rotate the scanned image if the text lines are not aligned horizontally. The rotation can be performed automatically or manually by moving one of the handles (two blue points) with the mouse, by typing in the rotation angle or by pressing the CTRL key (CTRL+SHIFT for bigger steps) and turning the mouse wheel.

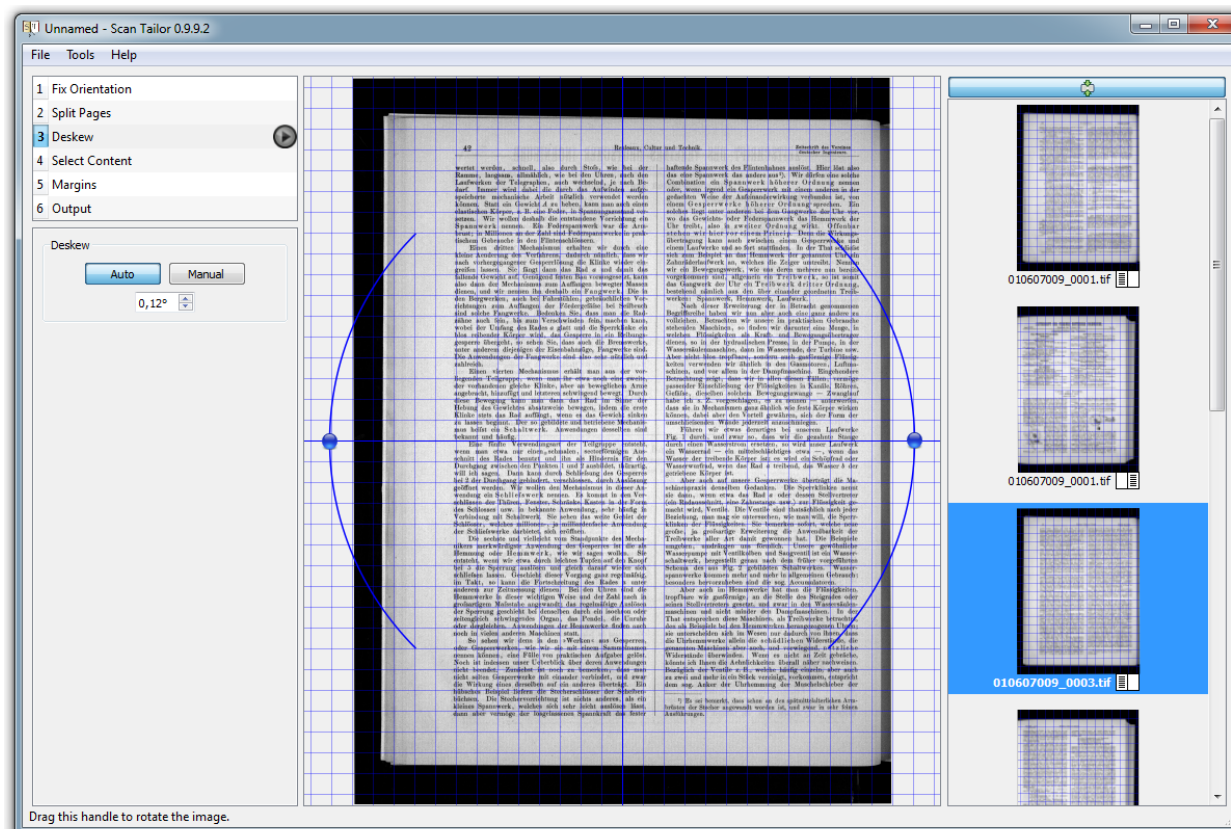



Figure 22 - ScanTailor - Result of automatic deskewing a page

Please follow the instruction:

- **Go to the first image** by clicking at the first thumbnail picture (In other case batch process runs from the selected page to the last page only.)
- Push the **play** button  to launch the batch process for all pages.
- **Check visually all pages** by browsing through all thumbnails to check up the result of automatic mode or use Page up/Page down keys. Criteria are text lines, borders of text blocks, lines in header or footer.
- If there are images where the pages rotation is **not corrected exactly**, turn the page **manually**.

Step 4 – Select Content

In this step the region of the page content (text, graphics, pictures etc.) has to be defined. The area outside the content box will be filled with white colour. That means every structure also text or page numbers etc. outside the blue rectangle will disappear in the later output page.

The step works also in automatic or manual mode. In a case of wrong automatic determination of the region you can change the position and the size of the content box by dragging the lines or the corners with the mouse.

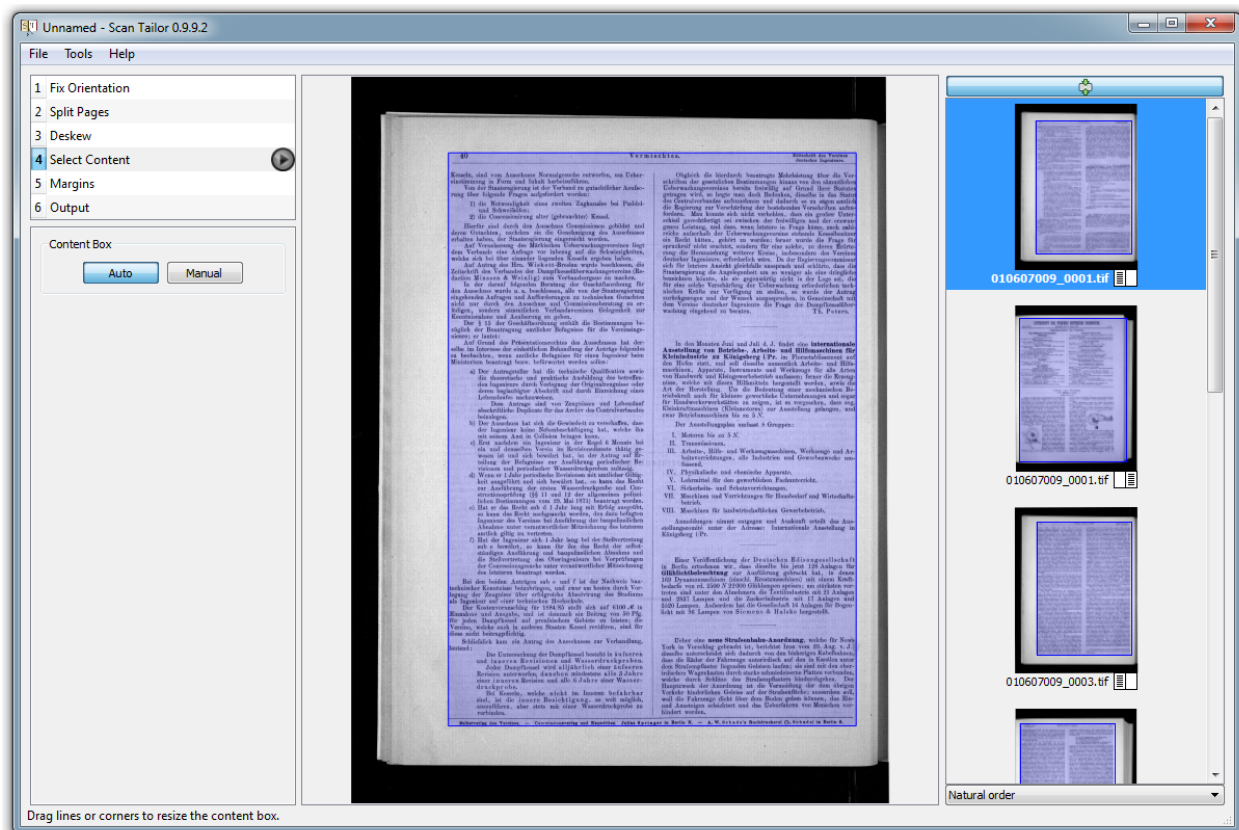



Figure 23 - ScanTailor - Automatically determined content box (blue rectangle)

Please follow the instruction:

- **Go to the first image** by clicking at the first thumbnail picture (In other case batch process runs from the selected page to the last page only.)
- Push the **Play** button  to launch the batch process for all pages
- **Check visually all pages** by browsing through all thumbnails to check up the result of automatic mode or use Page up/Page down keys. Observe that all text lines, footnotes, page numbers etc. are inside the content box. Otherwise these elements will disappear later in the output page.

Please note also that no stains (dirty marks) enlarge the regions needlessly. In extreme cases you should exclude these stains from the region, because in later steps all pages will be as wide or high as the largest content box plus margins.

Blank pages with no content usually cause an incorrect automatic selection of the content box. But in this case there are no manual changes necessary. If there are problems with visible blocks in the output image or problems in the step *Margins* you can go back to this step and you can remove the content boxes of the blank pages with the right mouse button

(Please note: Do not run automatic mode again after manual corrections. You will lose your corrections.)

Step 5 – Margins

This step defines the margins for the output pages. Please note that this step is a compromise between the visible quality of your produced output pages provided in the online portal and the efficiency of your work. Please do not waste too much time for this step because the *thinkMOTION* project has to produce many *Europeana* items. But we have also to deliver not the best possible but a reasonable quality.


It is normal that you need the same time for step 5 as for steps 1 to 4 together. But it is too long if you need three times as much. Usually you will need more time for the first books or articles but you will become faster over time.


The margins (in mm or inches) are set for: top, bottom, left and right. A value of the margins of normal text pages between 5 and 10 millimetres delivers a reasonable layout of the finished pages. Pages with large drawings, reaching into the original margins, can require a smaller margin setting. The margin area will be filled with white colour.

All pages of a source like a book, which are in the same size in the original, should have the same size in the finished pages. For this you have to choose the option **Match size with other pages**. So called fold out pages, which are normally larger than the majority of the pages, should be handled separately.

The alignment of the content in the page can also be set. The alignments should be set in this way: find an upper, lower or more seldom a left or right line, number etc. which is at the same position for the majority of the pages and set the alignment setting according to the alignment of this “object”. Normally, you have a header line or the text block begins at the same position at each page so the top alignment delivers mostly the best result. For pages with a special layout, e.g. on pages with the beginning of a new chapter the alignment of the text block can be moved down, a separate alignment could be necessary.

In a first run please follow the instruction:

- **Choose** a typical content page from the thumbnail list
- **Set the margins** (between 5 and 10 mm)
- Press **Apply To ...** button, select **All pages** and press **OK**
- Push the **Play** button  to launch the batch process for all pages
- **Check visually all pages** by browsing through all thumbnails to check up the result of automatic mode or use Page up/Page down keys.
- Do go on with the second run described below

Often the first and last few pages are nonstandard pages concerning the layout. Such pages are the title page, pages with library stamps at the margins, the table of contents pages, the index pages, blank pages, pages with very large drawings or especially the fold out pages. The margins of these pages should be similar to the margins of the majority of all pages. If one of these pages defines the overall size of the margins in an extreme manner you should first check the correct setting of the content box of this page in step 4. Sometimes stains can be responsible for too high or too wide content boxes. If the settings in step 4 are correct and you still have the effect, you should change the margins and alignments for this page separately in step 5 (use the *This page only in the Apply To ...* options). The default setting is that the left and right margin and also the top and bottom margin are changing together. You can loose this linking with the chain symbol .

If there are blank pages among the problem pages you should remove these blank content boxes in step 4, because blank pages should not define the size of all pages. Blank pages normally do not need any margin setups.

If the original page paper size or sheet size is larger than the size of the normal pages (e.g. fold out pages), you ought to uncheck the option *Match size with other pages* in this case. But do this for such pages only!

To find out the problem pages *ScanTailor* has a utility function. You can change the order of the thumbnails in the right sector in three ways:

- *Natural order* (the default setting)
- *Order by increasing height*
- *Order by increasing width*

If you choose the options *Order by increasing height/width* you can find the problem pages at the end of the thumbnail list.

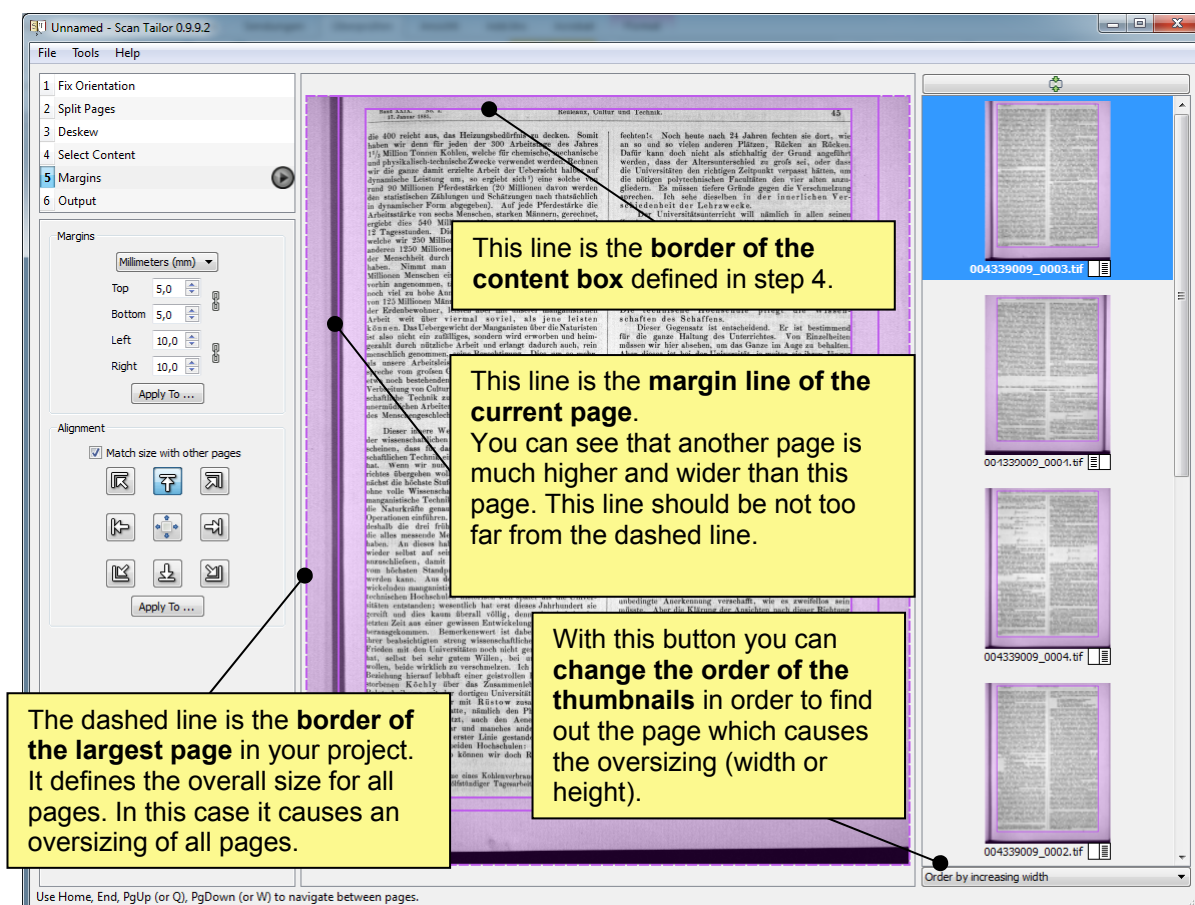


Figure 24 - Meanings of the lines in the margins setup and the changeable order of the thumbnail pictures

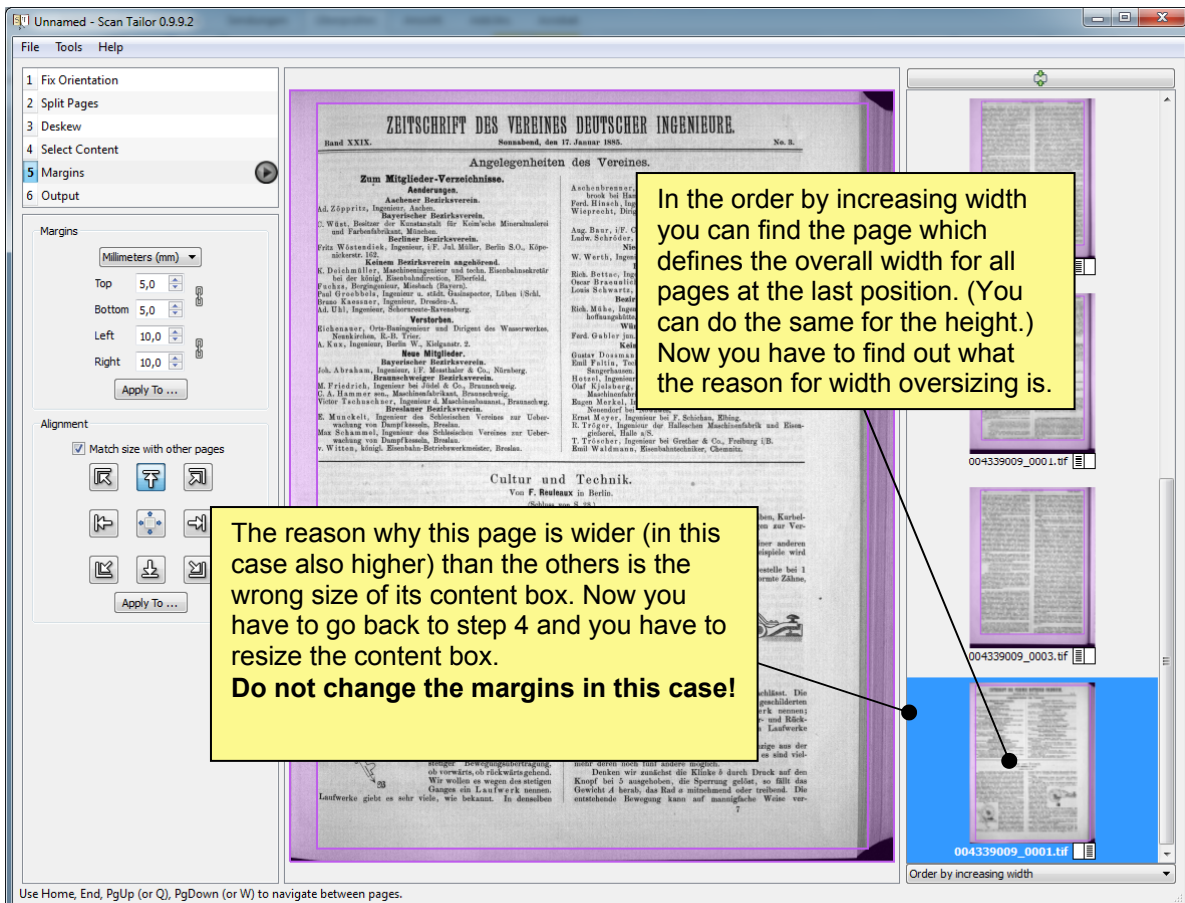


Figure 25 - Finding oversized pages by using Order by increasing height/width and reason for oversizing

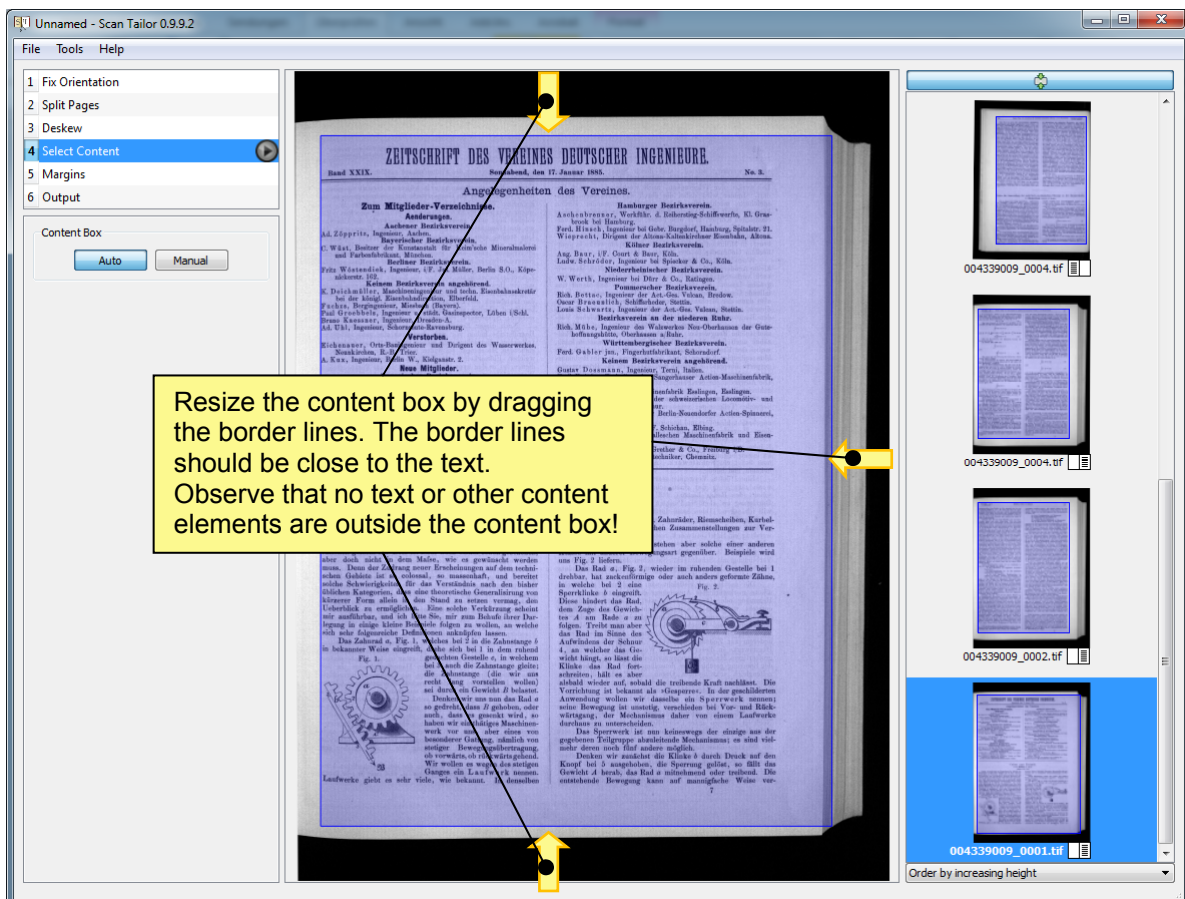


Figure 26 - Resizing the content box in step 4 as the reason for the oversized page in figure 25

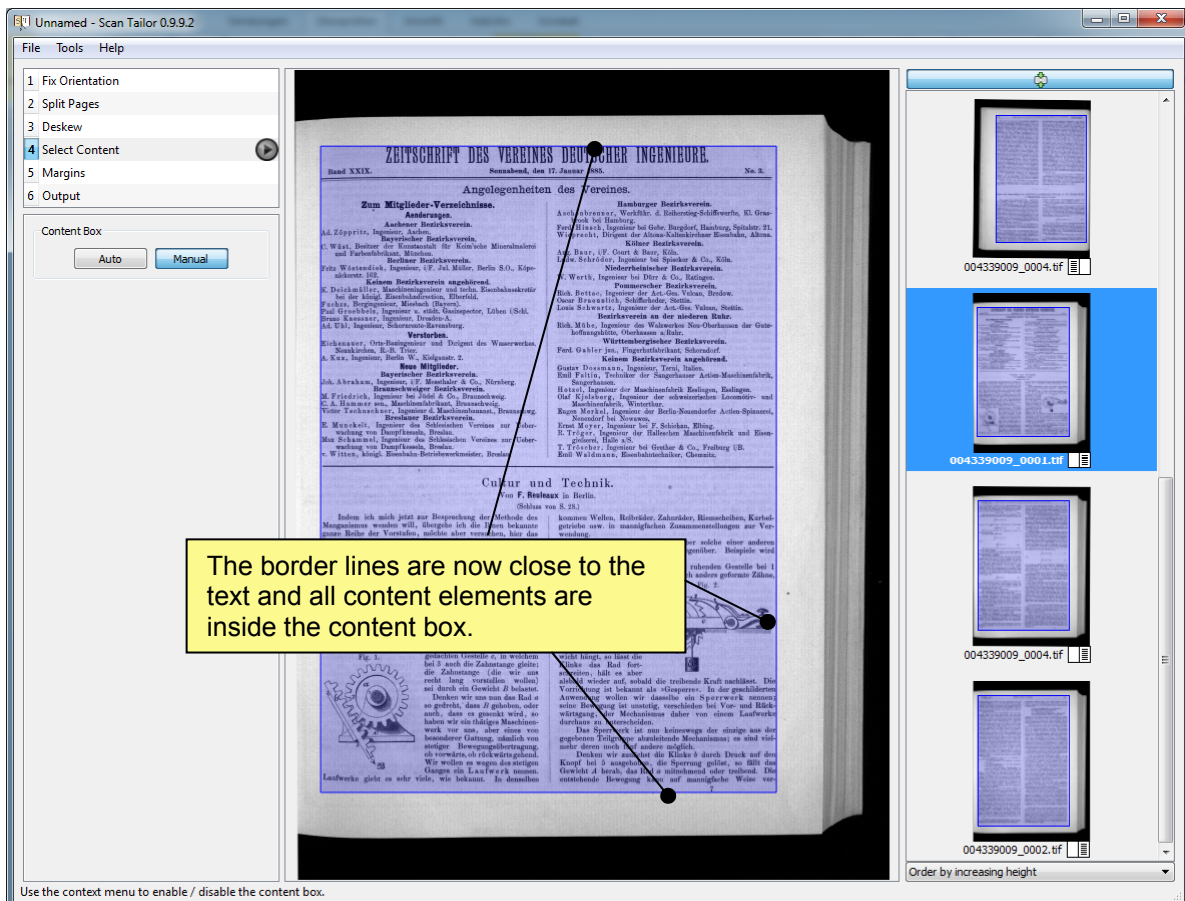


Figure 27 - Criteria for a well-defined content box

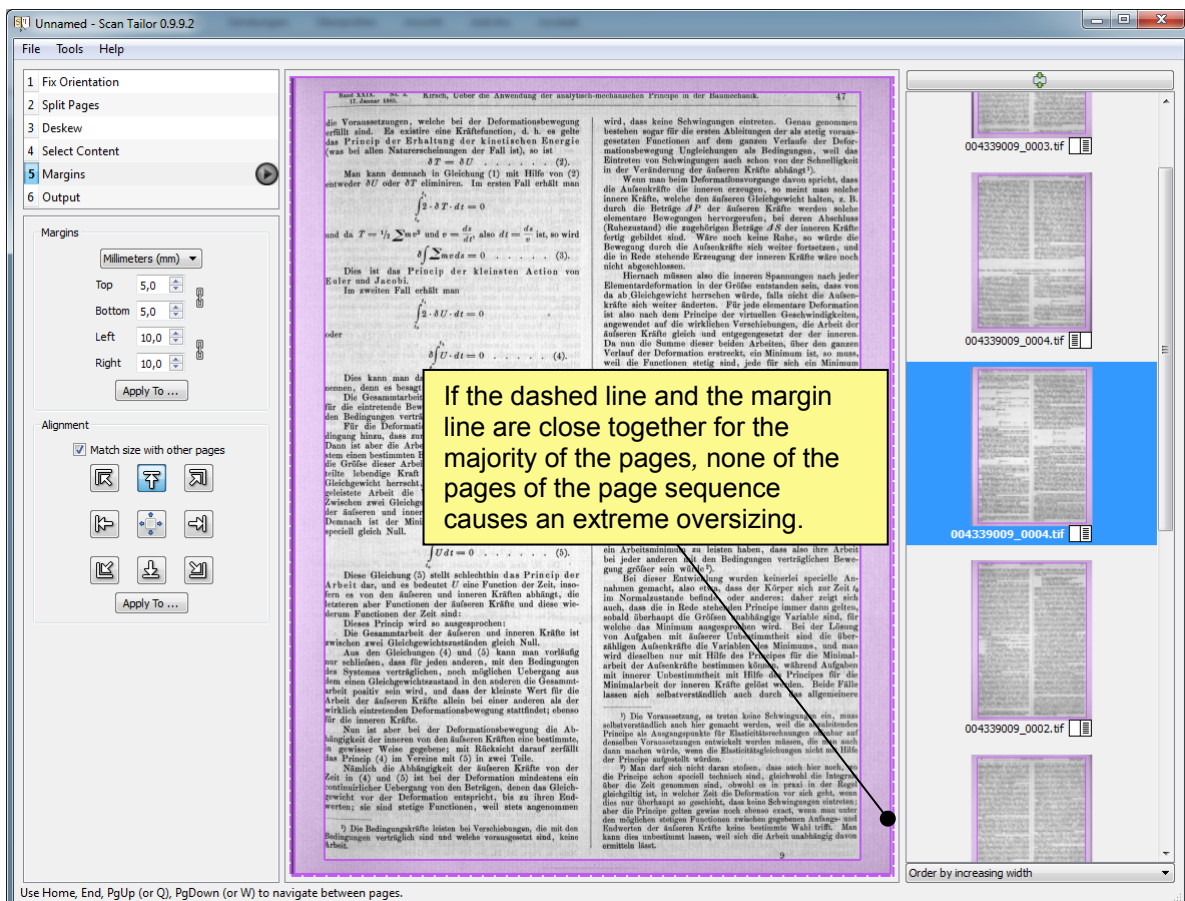


Figure 28 - Criteria for well-defined margins and not extremely oversized pages

In a second run please follow the instruction:

- Choose the option **Order by increasing height** at the bottom of the right sector and **scroll to the end** of the list.
- Check whether the **top and the bottom margins** of the last few pages in this order, which are responsible for the overall height of all pages, are set well.
- If necessary, change the settings of the **content box** (step 4) or the **margins** (step 5) as described above.
- Repeat both previous steps **check and change** until the result is reasonable (Please note: Make a compromise between quality and quantity!)
- Choose the option **Order by increasing width** at the bottom of the right sector and **scroll to the end** of the list.
- Check whether the **right and the left margins** of the last few pages in this order, which are responsible for the overall width of all pages, **are set well**.
- If necessary, change the settings of the **content box** (step 4) or the **margins** (step 5) as described above.
- Repeat both previous steps **check and change** until the result is reasonable.

Step 6 – Output

This step allows to set resolution, colour mode and despeckling (if necessary) of the output files.

The resolution of the output file can be established at fixed resolutions such as 300 dpi, 400 dpi, 600 dpi or at a custom value. Please **keep the resolution at its original value** to avoid quality losses! (Please note: Increasing the resolution does not lead to better quality!) You can find out the resolution with a picture viewer like *IrfanView* (see figure 29). In *IrfanView* choose *Menu bar/Image/Information*

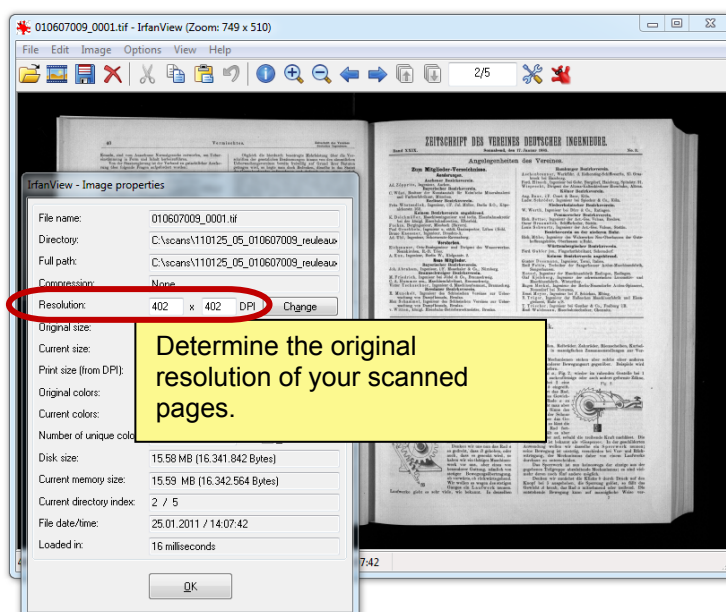


Figure 29 - Finding out the original resolution of a scanned image with picture viewer *IrfanView*

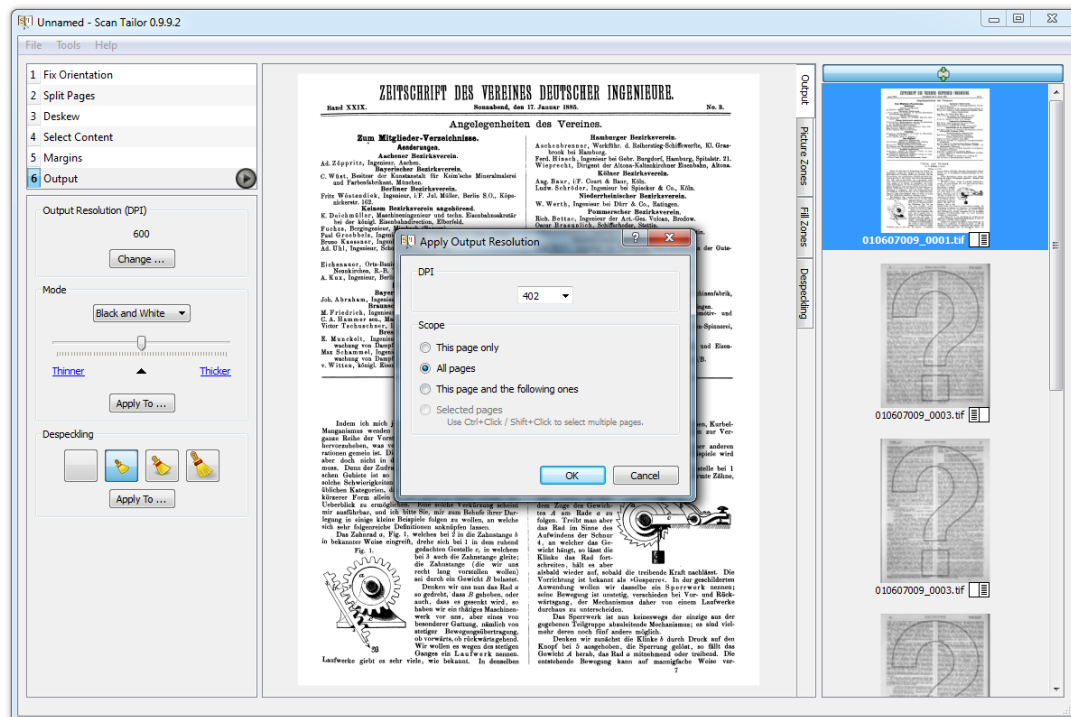


Figure 30 - Typing in the exact resolution of scanned images for generating the output page files and applying to all pages

In *ScanTailor* you can change out of three **colour modes**:

- *Black and White*
- *Color / Grayscale*
- *Mixed*

The quality of the resulting output images in the online portal do not only depend on the colour mode setting, but also on the type and the state of the original scanned pages (text contrast, page background, pages with photographs, colour or greyscale scans etc.)


It is not possible to define a general mode which delivers the best result in all cases. This setting needs some experiences, which you will get after processing some more sources. Table 4 and the examples in the annex shall help you at the beginning.

Table 4 - Typical types of sources and recommended *ScanTailor* settings as well as the hints for processing, checking results and fixing faults (see also annex)

Type of source	ScanTailor settings with possible problems, to be checked and to be fixed			
	black/white	colour/grey scale	Mixed	Combination (each page in an individual mode depending on source type - use options in the associated previous rows)
Pages all-over imprinted (with background pictures or colours, e.g. magazines or brochures)	<ul style="list-style-type: none"> Not usable because loss of background picture or colour (Figure B4 in the annex). 	<ul style="list-style-type: none"> Do not use option <i>Equalize Illumination</i> (Figure B5 in the annex, colour change), In most cases margins in step 4 should be set to zero (white margins are not pretty) 	<ul style="list-style-type: none"> Not useable because colour gaps in full coloured areas with overlaid text (result can be blocks with black or white text and white or black background inside a coloured background picture – looks awful) 	<ul style="list-style-type: none"> Combination not usable because loss of colour (see rows for black/white and mixed mode)
Colour print with white background and photographs, texts and/or photographs and/or drawings are coloured	<ul style="list-style-type: none"> Not usable because loss of colour 	<ul style="list-style-type: none"> Use options: <ul style="list-style-type: none"> <i>White margins</i> <i>Equalize Illumination</i> (check the colour quality for colour prints, and switch it on or off) Quality depends on type of source, paper quality, contrast etc. Result can have more or less faults, which cannot be fixed with <i>ScanTailor</i>, such as: <ul style="list-style-type: none"> Noisy background Different background grey scale and margin whiteness Disturbing rear-side-shining-through-effect (Figure B3, Figure B7, Figure B9 in the annex) 	<ul style="list-style-type: none"> Good quality Texts are in black and white Photographs and drawings are in colour Problems: <ul style="list-style-type: none"> Picture regions have to be resized in some cases manually (Figure B11 in the annex) Only usable if text is black (otherwise loss of text colour) 	<ul style="list-style-type: none"> Good quality No combination with black/white mode Combination between <i>colour/grey scale</i> and <i>mixed</i> possible if conditions are OK Step A: choose mode for pages which are in the majority for batch setting to all pages Step B: Switch pages which are in the minority into the other mode page by page and do corrections if necessary Problem: Only efficient for sources with an unbalanced number of types of pages
Black and white print with grey scale photographs and /or drawings	<ul style="list-style-type: none"> Not usable because loss of grey scale information in the photographs or drawings, because images get only two colours, black and white (Figure B12). 		<ul style="list-style-type: none"> Only recommended for sources with a lot of photographs, else use combination of black/white with mixed, see right table cell Texts are in black and white Photographs are in colour/grey scale Problem: Picture and drawing regions have to be resized in more or less cases manually (Figure B11). 	<ul style="list-style-type: none"> Only efficient for sources with a few photographs, else use mixed mode Step A: Set all pages in black/white Step B: Switch pages with photographs into mixed mode page by page and resize picture regions if necessary
Black and white print with text and line drawings but without photographs, grey scale or colour (most of the technical books)	<ul style="list-style-type: none"> Less rear-side-shining-through-effect (Figure B7) Homogeneous white background and margins Problem: smoothed types with a little more unpleasant readability, bad shading resolution in the drawing (Figure B10). 		<ul style="list-style-type: none"> Not useful because there are no grey scale or colour pictures Problem: Picture and drawing regions have to be resized in more or less cases manually (Figure B11). Better use black and white or grey scale mode 	<ul style="list-style-type: none"> Not useful because there are no grey scale or colour pictures Better use black and white or grey scale mode

 Bad settings. Do not use them!

 Settings can be used for *thinkMOTION* project. A careful quality check-up and perhaps a more or less expensive fixing process are necessary.

 Recommended *ScanTailor* settings for *thinkMOTION* project

Please give us a feedback if you have made other experiences!

The despeckling function allows cleaning the output file in a range of four levels. This choice is active only for the text regions of the *mixed* mode and for the *black and white* mode.

You can apply this setting to the current page or automatically to all pages. Please note: When using the automatic mode, it is recommended to check visually all pages to guarantee that none of the content has disappeared in the output files. For this you can use the *Despeckling* tab, described below. To work efficiently you should use this function only in special cases.

At the right edge of the main window you can find four tabs *Output*, *Picture Zones*, *Fill Zones* and *Despeckling*.


The **Output** tab shows the state of the page for the next workflow step after *ScanTailor* or with other words the result of your work of the steps 1 to 6.

Picture Zones, which is only available in *Mixed* mode, shows you the regions of the pictures in the current page, which are set to grey scale or colour mode, and the text regions, which are set to black and white mode. In some cases it is necessary to change the regions for a better visual quality of the output pages (see figure b11 in the annex).

Fill Zones is normally used for removing dirt stains in the page. You should use it to hide the end of the previous or the beginning of the next article if we are not allowed to show the neighbour articles. The reason for this can be the missing rights of use for the other articles or the lack of relevance of the neighbour articles. In this case you can draw a polygonal region over the part of the article without the rights of use and fill it with the background colour by using the right mouse key menu.

Despeckling shows the regions where your despeckling settings have changed the page content. It can be used to check whether important parts are removed.

Please follow the instruction:

- **Find out** the original **resolution** with a tool like *IrfanView* (see figure 29)
- **Press** the **Change...** button, select custom resolution and type in the **original resolution** (e.g. resolution is 402 dpi – type in exactly 402 dpi not 400 dpi!)
- Set the **scope** to **All pages** and press **OK** (Premise: All pages are scanned with the same resolution, which should be the normal case.)
- **Find** your **type of source** in the **Table 4** above and **use** the **settings** for the colour mode **of the green or yellow coloured field** (Attend to settings of the *White margins and Equalize Illumination*!)
- Choose **Apply to... All pages** for the colour mode and press **OK**
- Push the **Play** button  to launch the batch process for all pages
- If you have chosen a combination of more than one colour modes from the table, change the colour mode settings for a minority of the pages individually.
- Use the **Fill Zones** tab and draw a region to **hide articles** with missing rights of use if necessary. Use the right mouse key to fill the region with the background colour.
- **Save your ScanTailor project** file into the folder `application_data\scantailor\` and name it with the 9-digit DMG-Lib ID (xxxxxx009) ending with “009”

5. Optical Character Recognition (OCR) of Scanned, Prepared and Cleaned Documents

5.1 Overview

After cleaning and preparing the scanned documents with *ScanTailor* all sources containing printed text have to be processed by an optical character recognition tool. For the *thinkMOTION* project we recommend the Software **ABBYY FineReader 10 Professional**.

ABBYY FineReader converts text images into machine-readable (or editable) text. Before performing OCR, the program analyses the structure of the entire document and detects the areas that are containing text, barcodes, images and tables.

5.2 Download and Installation

First of all you have to buy and to install the software. For installation follow the instructions of the setup program.

5.3 Processing Steps for OCR

After installing *FineReader* successfully, do the following steps.

Step 1: Start the program *ABBYY FineReader*

Start *ABBYY FineReader* by using the icon  at your desktop or chose it from the Start/Program button at your Windows task bar.

The program window opens and you have the choice between different output formats and different tasks (figure 31). Although it seems to be the correct task for the *thinkMOTION* project, do not use the task *Adobe PDF => Convert to Searchable PDF Document* because it saves your file as PDF format instead of the project recommended PDF/A format. Please uncheck the option *Show at startup* so that this window does not appear at the next start. (You can launch this window at any time by pressing the *New Task* button in the menu bar.)

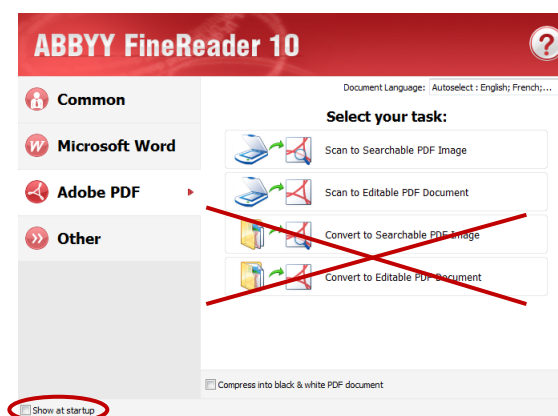


Figure 31 - *FineReader* task window after startup – do not use it!

Please follow the instruction:

- Ignore the start-up window with the tasks if it still appears.
- Choose from the menu **File** the option **New FineReader Document**.

Step 2: Import the cleaned and prepared pages into the *FineReader* project

In this step you have to import the single page images, which are cleaned and processed by *ScanTailor* Software, into your *FineReader* project.

Please follow the instruction:

- Choose from the menu **File** the option **Open PDF File/Image...**
- Navigate to the **ScanTailor output directory** (e.g. C:\scans\110125_05_010607009_reuleaux_cultur_und_technik\scans\out) and **choose all pages** (select the first TIFF file, hold down the SHIFT key, select the last TIFF file).

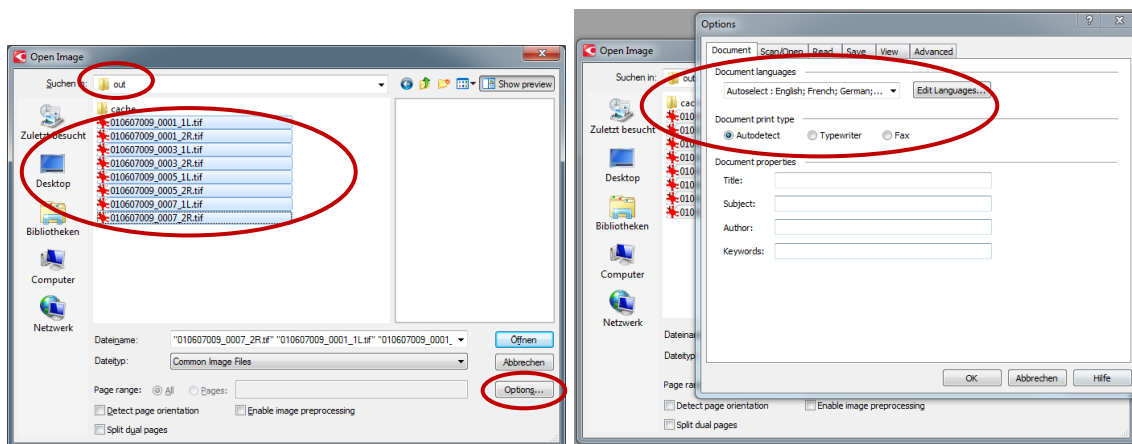


Figure 32 – *FineReader* – project configuration

- Press the **Options...** button and check or change the settings at the tab **Document**
- Set the **Document languages** to **Autoselect** for normal use.
- Set the **Document print type** to **Autodetect** for normal documents or to **Typewriter** if you have a document written in typewriter font.
- You can leave the **Document properties** blank – it will be ignored in later steps anyway.
- Change to tab **Scan/Open**

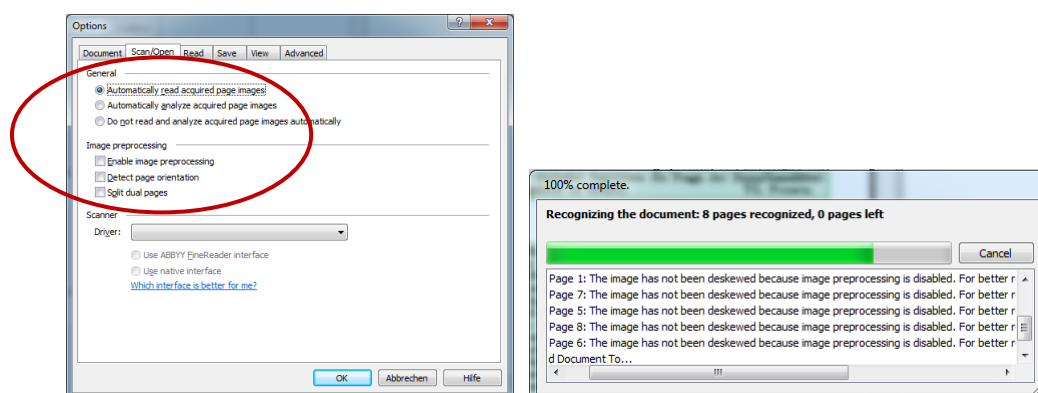


Figure 33 – *FineReader* – project configuration and progress of reading and analysing the pages

- Use setting **Automatically read acquired page images**.

- Make **no Image processing** (We have it already done in *ScanTailor* – a second time needs a further check-up and manual correction process!). Uncheck all options for this.
- Press **OK** to close the options window and then **Open** in the Open Image Dialog.

The program recognises all the pages of the project one by one. This step can take some time. The displayed window shows you the progress.

Step 3: Work with *FineReader* document

After the reading and the analysing process all opened files can be previewed on the left side of the window. The active area is split into two windows, containing the original *Image* and the *Text*, optically recognised and converted into machine-readable text.

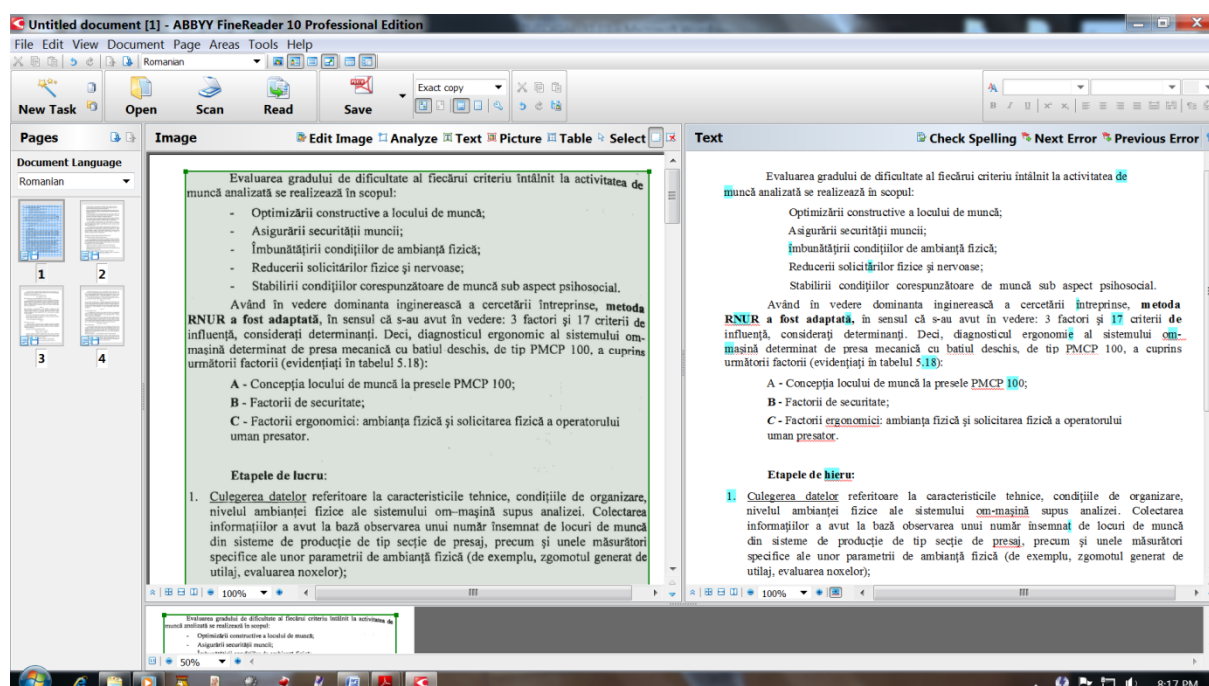


Figure 34 – *FineReader* – OCR result with highlighted unsurely recognised characters

The first page is automatically displayed in these two windows. In the *Text* window (right) you can see the characters highlighted, which are not surely recognised, and you can correct them. If there are too many faults, please check your setups for *language* and *print type*.

Usually most of the highlighted characters are still correct and the user cannot see wrong characters in the online portal anyway. The only disadvantage is that these incorrect words are not found in some cases by the full text fuzzy search. But important information of the source such as the title, the author, the publisher and the issue date are entered as meta data in the database manually and will be found by the user search in all cases anyway. For that reason and to work efficiently in the *thinkMOTION* project it is not recommended to apply any corrections to the OCR result.

Step 4: Exporting the *FineReader* document into a PDF/A document

The next step is saving the document as a PDF/A file (PDF/A is a standardised PDF format for archiving). Do not use the normal PDF format, which is also offered.

Please follow the instruction:

- From the menu **File** choose the option **Save Document As...**, choose **PDF File/A Document** (see figure 35)

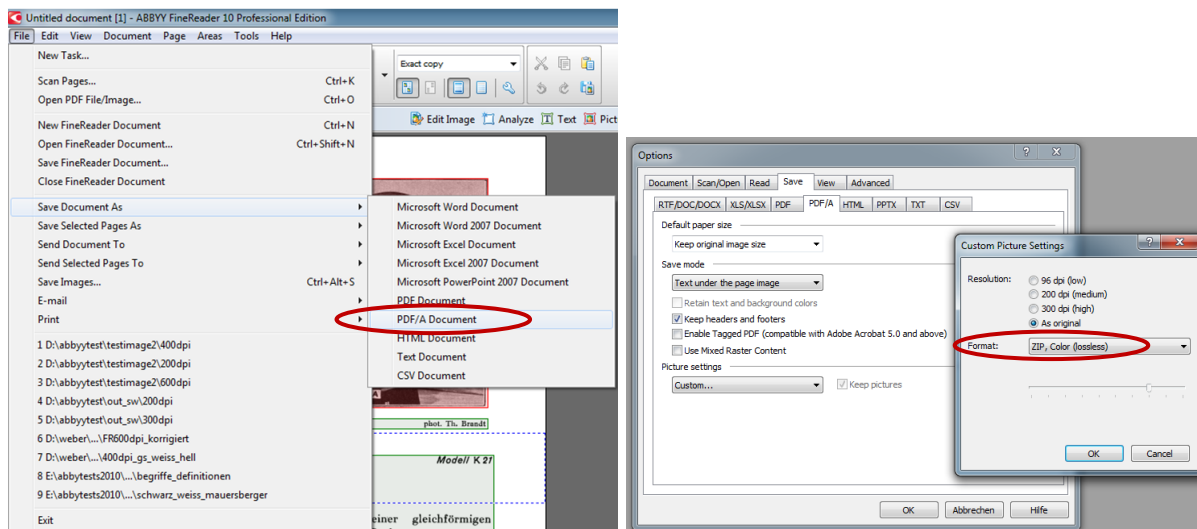


Figure 35 – FineReader – configurations for export as PDF/A file

- Navigate** to the **PDF** folder in your scan folder (e.g. C:\scans\110125_05_010607009_reuleaux_cultur_und_technik\pdf) **Do not save it yet!**
- Press the **Options...** button and **check or change the settings** at the tab **Save PDF/A**
- Set **Paper size** to **Keep original image size**
- Set **Save mode** to **Text under the page image** and uncheck all options, only Keep headers and footers can be active.
- Set the **Picture settings** to **Custom...** and choose the **Resolution As original** and set the **Format** to **ZIP, Colour (lossless)**.
- Press **OK** twice
- In the save dialog window use the **dmglID**, that ends with "009", as **name** of the PDF file and press the **Save** button.

Then the program *Adobe Reader* opens the file containing a number of pages equal to the number of images. Please check the quality of the page layout, of the text and of the pictures and drawings.

Please note: The selected options in all tabs of the *FineReader* program are saved and should represent the default mode on further use of the program. But you should better check them again for the next time or if other persons have been working with your PC.

6. Data Handling – Part II: Upload the Folder Structure of a Processed Document

After scanning and processing a document, all the scanned raw data, the project file and PDF/A file containing OCR output are in the according folder structure. It is then necessary to upload this structure and the files to the server. To achieve this, please follow these steps.

Please follow the instruction:

- In the ProDB **press Scan** to navigate to the *Scan Workflow Overview*
- Find the table entry for the document you would like to upload. **Check** that its workflow state is **Reserved for scanning**. (Figure 7, Label 1).
- **Press the Upload button** in the *actions* column for the document (Figure 36). A pop-up will ask you if all necessary content (e.g. raw scans, PDF file) has been placed into the correct subfolders of the predefined folder structure (Figure 37). **Press OK** to continue.

DMG LIB

Search in GBV Search in TUI library Contact data of TM Partners Help for proDB handling Deutsch

show input helper (Ctrl + Alt + T)
03.03.2011 11:47:00 You are logged in as: reessing

?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI ?Task ?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI ?TestM #List #Str. #Rgt. Scan Upload #Imp. Logout #Ws Adm.

Scan Workflow Overview

Show documents changed during the last 31 days Refresh

Showing documents changed between Mon Jan 31 00:00:00 CET 2011 and Thu Mar 03 11:47:00 CET 2011

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	Reserved for scanning.	110303_03_004339009_Brown_A_Collection_of_Mechanical_Movements	Exists on local harddisk.	Scan document and Upload content.
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	Not scanned.		(N/A)	Reserve for scanning.
10505009	Reeßing Test Document for Scanning and Uploading	Not scanned.		(N/A)	Reserve for scanning.

Local disk access

This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk.

Local content folder:C:\scans [Change folder](#)

Figure 36 - Button for uploading the folder structure of a document

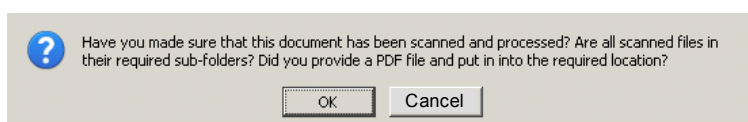


Figure 37 - Pop-up that asks if all content is in place and ready for upload

- An Uploader window will appear and will list the folder selected for upload. Please check if this is the folder you intend to upload and press the **Next** button (Figure 39). Note: Your web-browser may display a security warning since the Uploader needs to access your local hard drive (Figure 38). Please select **Run**.
- The Upload window will display an indicator that reflects the progress of the upload (Figure 40). As soon as the transfer completes, you can press the **Close** button of the Uploader.
- If the upload was successful, in the ProDB the **workflow state** of the document will change to **Uploaded**. (Figure 41). Note: In some web-browsers it is necessary to press the **Refresh** button to see the new workflow state. If the upload did not complete successfully, an error window will display a log of the failed upload process.

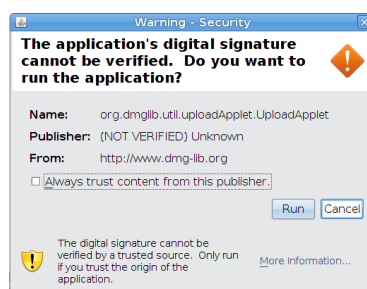


Figure 38 - Pop-up that asks permission to access the local hard drive in order to upload scanned content

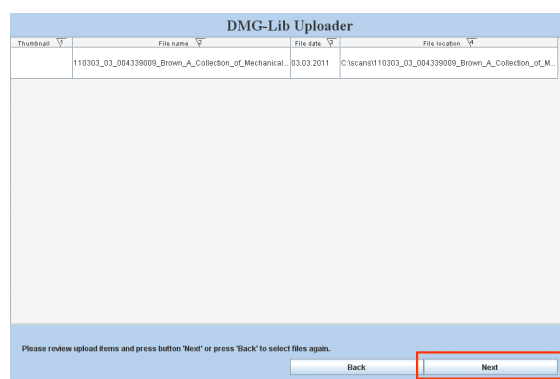


Figure 39 - Main window of the Uploader with the list of items scheduled for upload

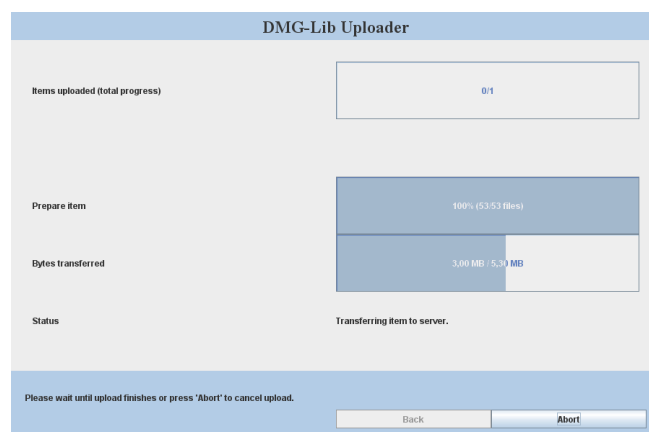


Figure 40 - Main window of the Uploader with a progress indicator

DMG LIB

Search in GBV Search in TUI library Contact data of TM Partners Help for proDB handling Deutsch

show input helper (Ctrl + Alt + T) 03.03.2011 11:50:25 You are logged in as: reessing

?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI ?Task ?Doc. ?Per. ?Mec. ?Img. ?AIS ?Col. ?URI #TestM #List #Str. #Rgt. Scan Upload #Imp. Logout W Adm.

Scan Workflow Overview

Show documents changed during the last 31 days Refresh

Showing documents changed between Mon Jan 31 00:00:00 CET 2011 and Thu Mar 03 11:50:25 CET 2011

id	author title	workflow state	scan folder name	local path existence	actions
4339009	Brown A Collection of Mechanical Movements	Uploaded.	110303_03_004339009_Brown_A_Collection_of_Mechanical_Movements	Exists on local harddisk.	
5720009	Kerle Kurt Hain, in: Distinguished figures in mechanism and machine science	Not scanned.		(N/A)	Reserve for scanning.
10505009	Reeßing Test Document for Scanning and Uploading	Not scanned.		(N/A)	Reserve for scanning.

Local disk access

This web-page needs to access your local harddisk to create scan folder structures and upload scanned documents. For this purpose, the web-page uses a Java(tm) applet which is represented by the "Local content folder" line below. The applet uses a path on your local harddisk, where it creates scan folders if you check-out a document for scanning. You may change the location of this folder by pressing the "Change folder" button below. Note: Changing the location of the folder will only tell the applet where to place scan folders. It will not move any existing data on your harddisk.

Local content folder: C:\scans

Change folder

Figure 41 - Successfully uploaded document

Annex

The following examples should give you a feeling for the problems and criteria, which have to be observed and used during the workflow of digitizing documents for the *thinkMOTION* project. The annex is divided into two parts. Annex A shows examples for working with *ScanTailor* and Annex B shows examples for working with *ABBY FineReader*.

Annex A – Check parameters for scanned files

This annex shows you some examples for bad parameters for scanning.

Please check by samples the compression parameters, the resolution and the number of colours in the image properties of your scanned files. You can use a picture viewer like *IrfanView* (Menu bar/Image/Information).

Figure A1 shows wrong compression parameter. In this case the scanned pictures are JPEG compressed inside a TIFF file.

Figure A2 shows the compression errors for the wrong parameters from Figure A1. JPEG compression is bad for character recognition and also the *ScanTailor* software is not able to import such files. Do not use it!

Please observe that the file extension do not tell you in any case how the file is compressed! TIFF-files for example can be inside lossy JPEG compressed (example_2_colour.tif in Figure A1). So it is absolutely necessary to check the file properties with a tool like *IrfanView*.

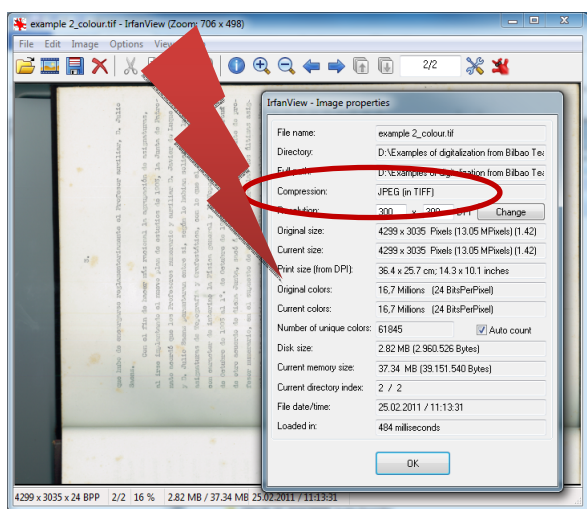


Figure A1 - Wrong compression parameter - JPEG compression in a TIFF file

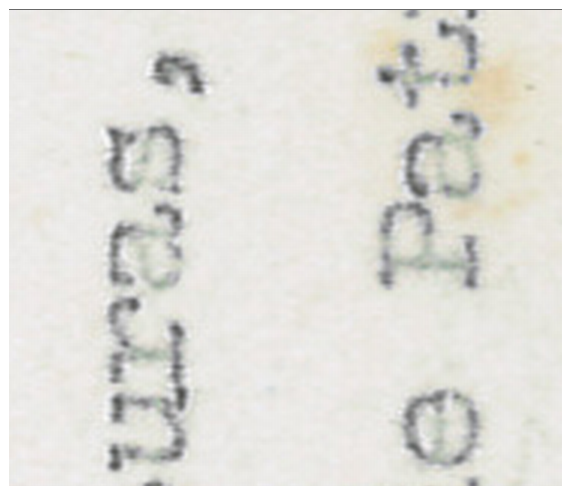


Figure A2 - JPEG compression error - blowing types (and lines) and JPEG boxes

Figure A3 shows another wrong compression parameter. In this case a Fax Encoding compression is used. This compression uses only two colours black and white. In Figure A4 you can see gaps in the types as the result of the encoding compression errors. This effect is extra strong on typewriter sources. Such compression errors are bad for character recognition. Do not use it!

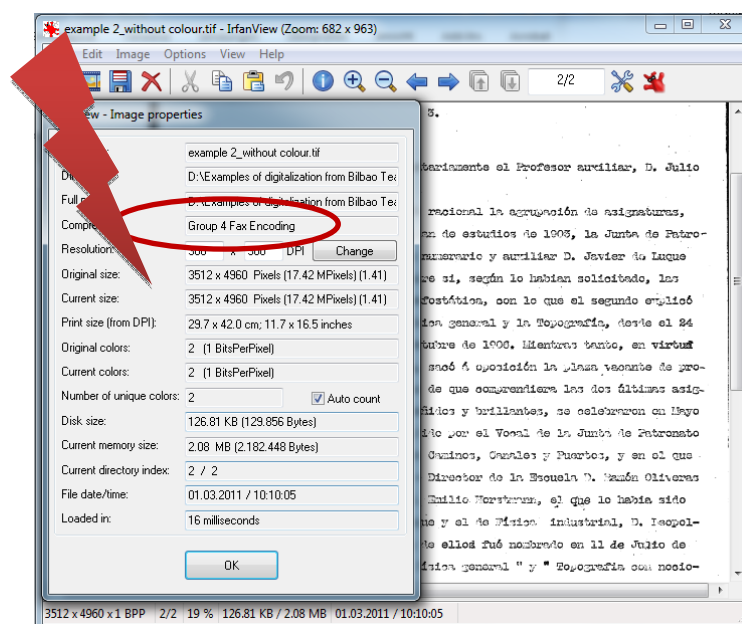


Figure A3 - Wrong Compression – Fax Encoding, 2 colours

Una Junta
al supues

Figure A4 - Fax Encoding compression error – gaps in the types, especially on typewriter sources

Annex B – Examples for working with ScanTailor

The following examples explain the influence of the settings on the quality of the output pages. Some settings are to be seen in more than one example but with different sources to make you sensitive for the problems, which you have to pay attention to when using *ScanTailor*.

B1 Influence of the colour mode settings for the colour mode *Grayscale* on the page and margin appearance

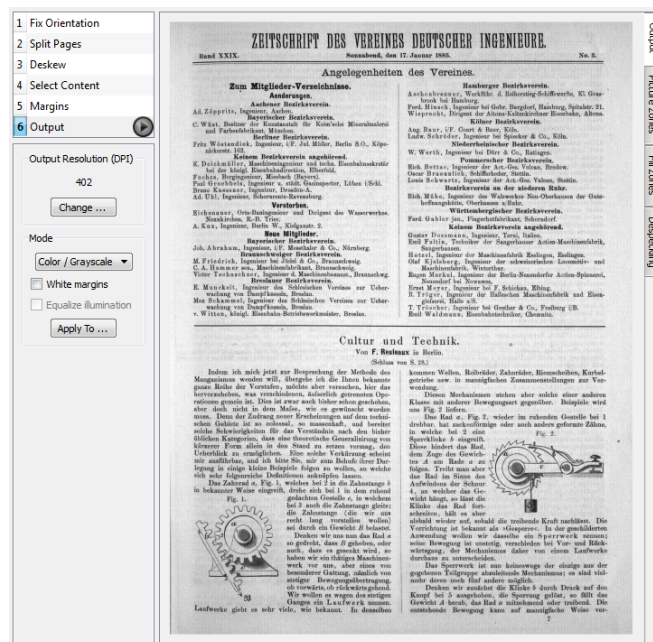


Figure B1 - Greyscale mode, margins are of the original scanned background colour

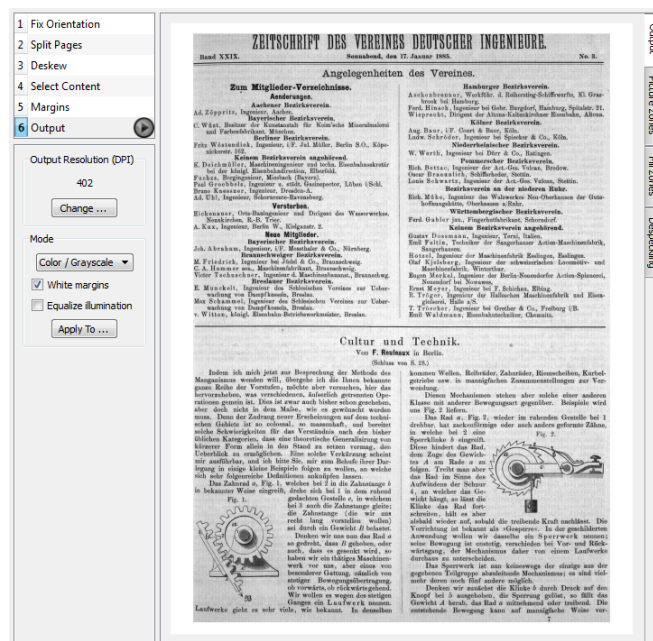


Figure B2 - Greyscale mode, margins are filled with white colour

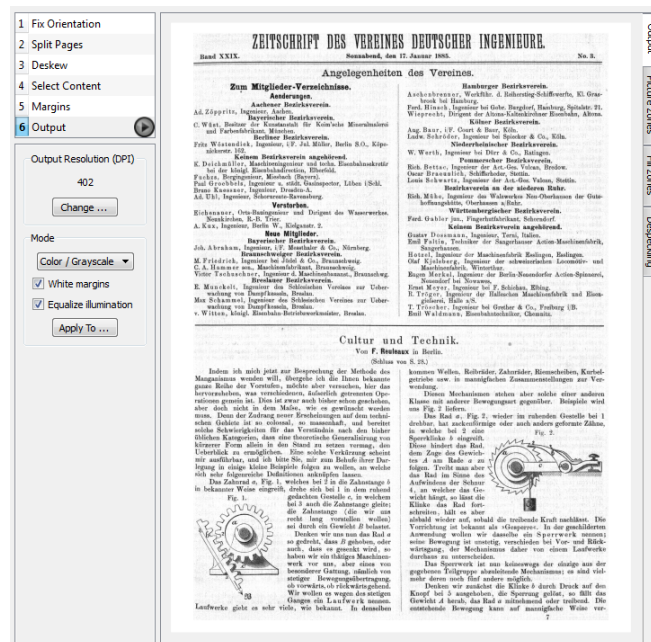


Figure B3 - Grayscale mode, margins are white filled, the background of the content region is equalized to the white margins, but the rear side content is visible again, even though with a lower visibility

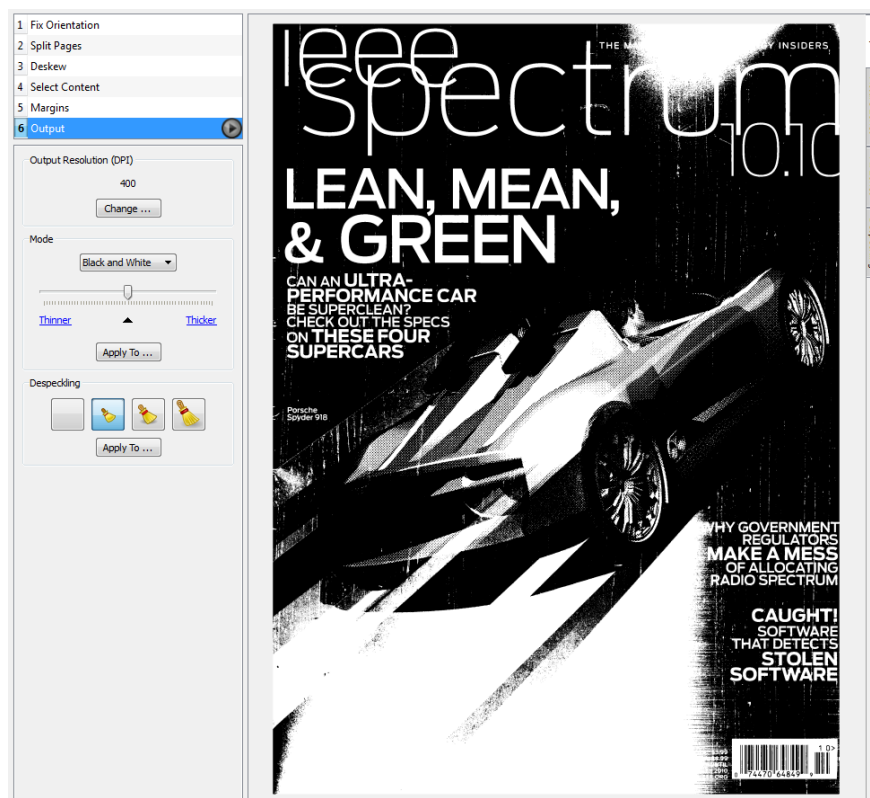


Figure B4 - Colour mode settings: **Black and White** mode. The colour is lost. Do not use this setting for all-over imprinted sources!

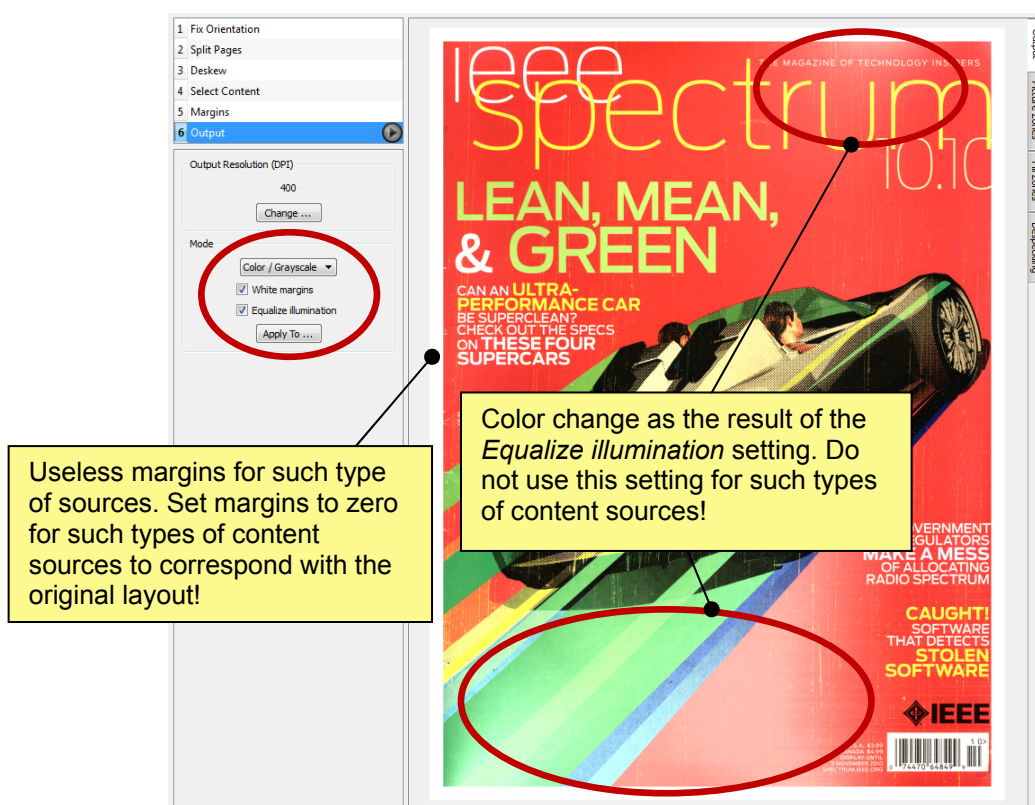


Figure B5 - All-over imprinted source type with 5 mm margin in each direction; colour mode settings see figure. Do not use the *Equalize illumination* setting for such types of sources!

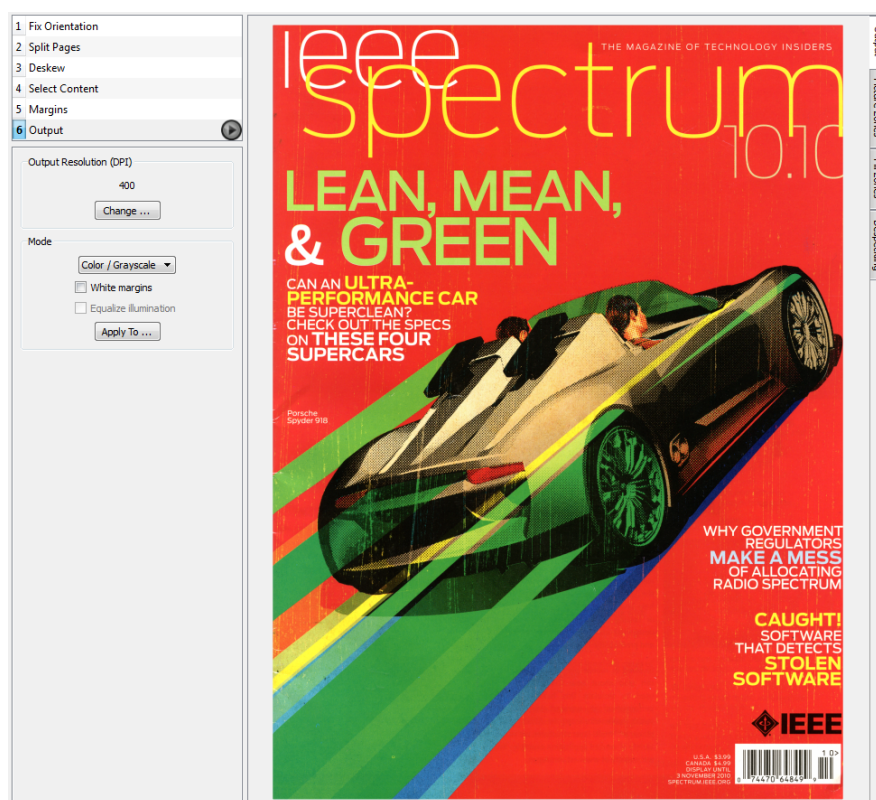


Figure B6 - Colour mode settings: *Color/Grayscale* mode; no *White margins*, no *Equalize illumination*, margins are set to zero. Colour and margin layout corresponds with the original. These settings are recommended for all-over printings.

B2 Influence of the colour mode settings on the text and drawing appearance

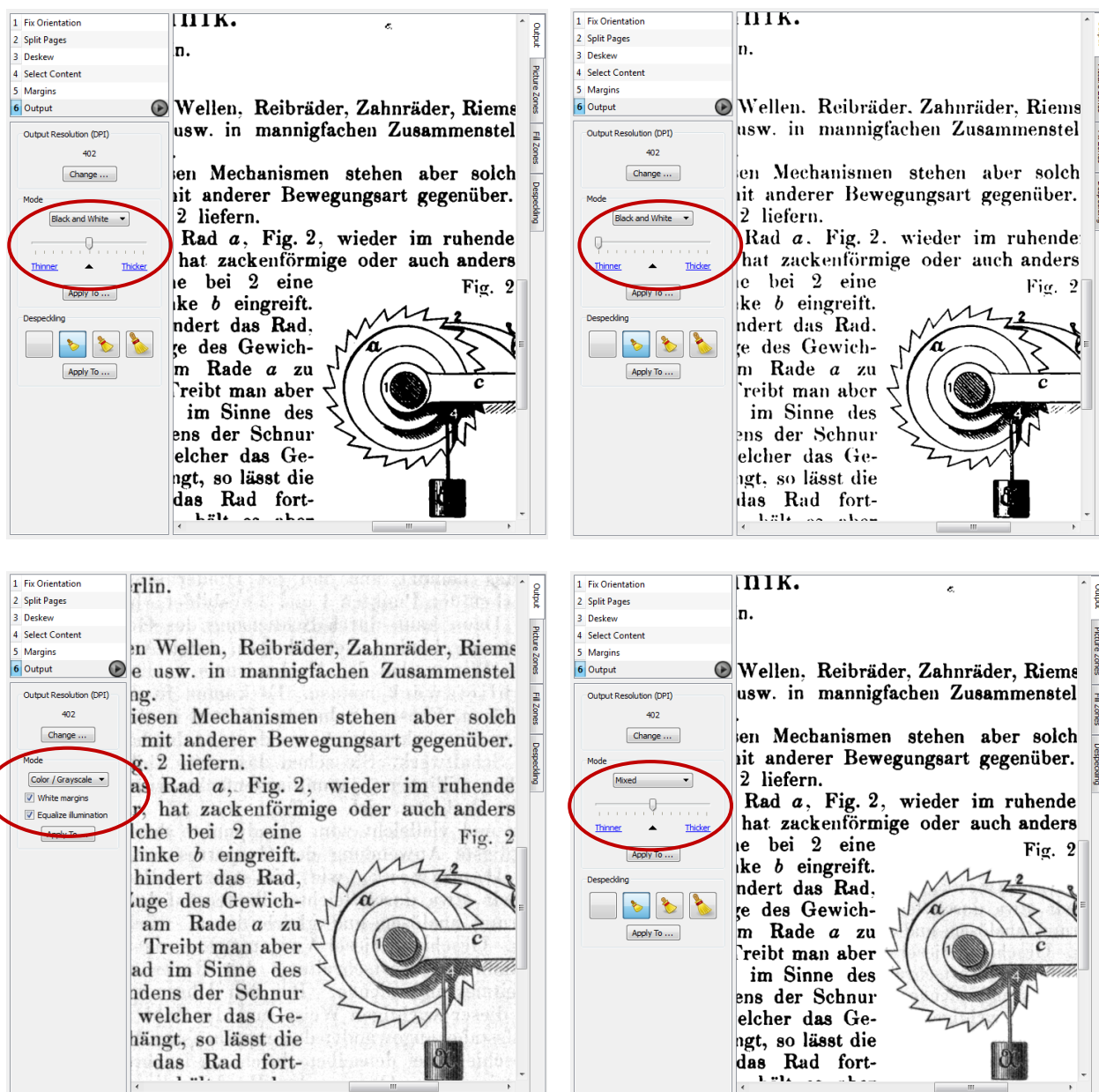


Figure B7 - Different settings for colour mode (right side of each figure) and the influence on the image - observe the "salt&pepper" at the text background, the rear-side-shining-through-effect and the letter written on the weight and the shading of the shaft in the drawing

B3 Influence of the colour mode settings on the picture/drawing quality

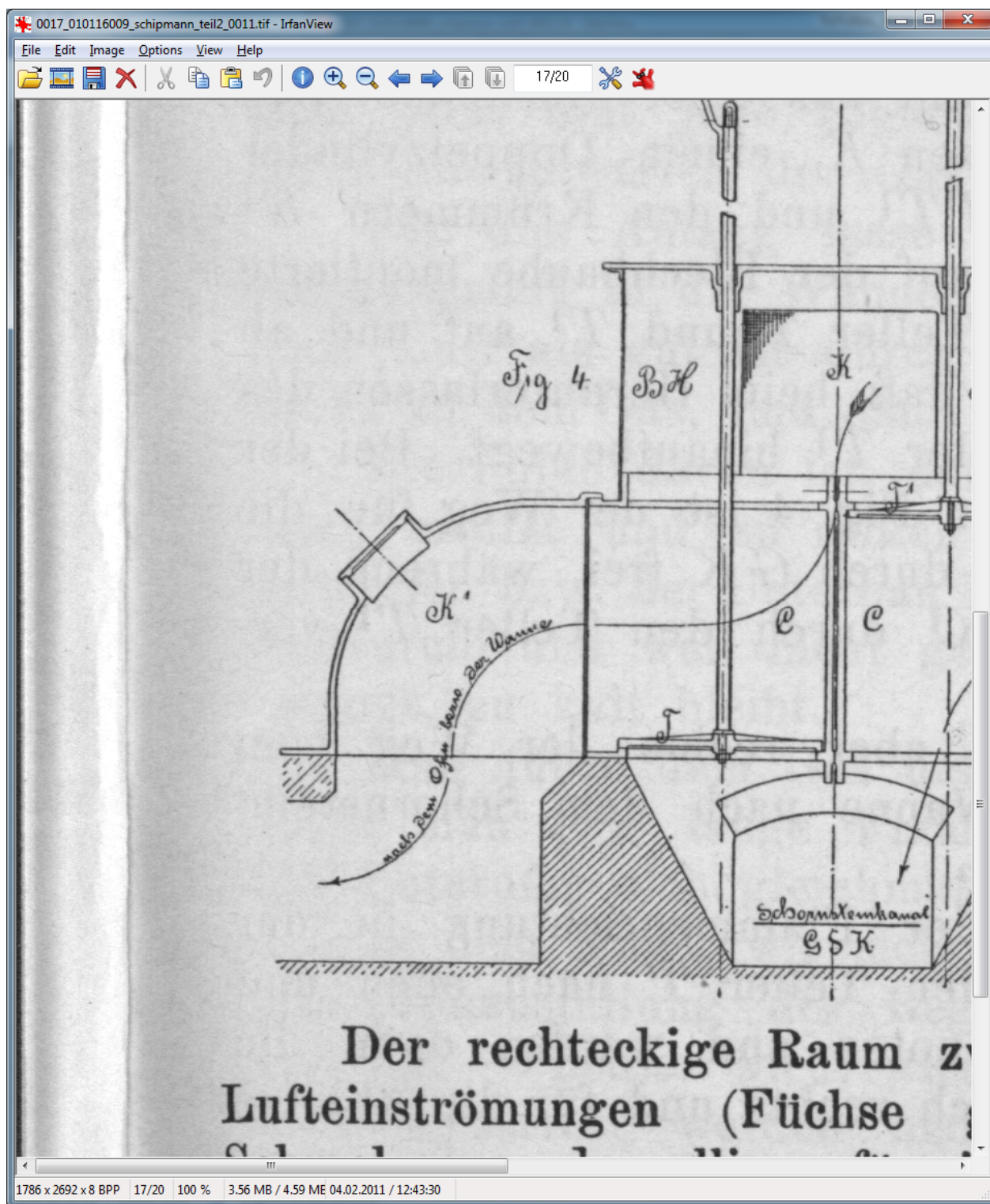


Figure B8 - Scanned image

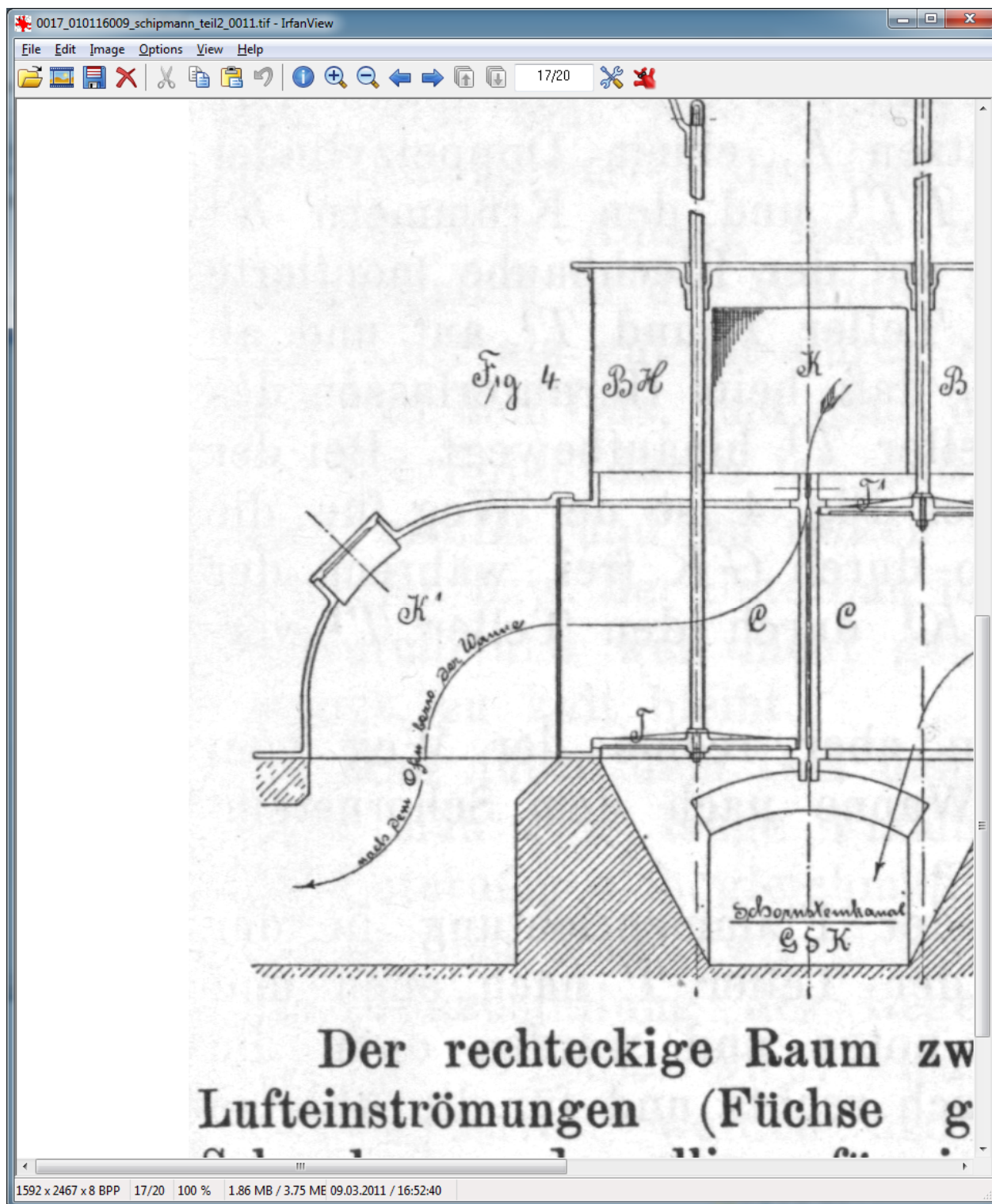


Figure B9 - *ScanTailor* processed image with settings: Grayscale, White Margins, Equalize Illumination
Advantages: better text quality, good readability of the handwriting in the drawing
Disadvantages: noisy background, rear side is shining through, large file size

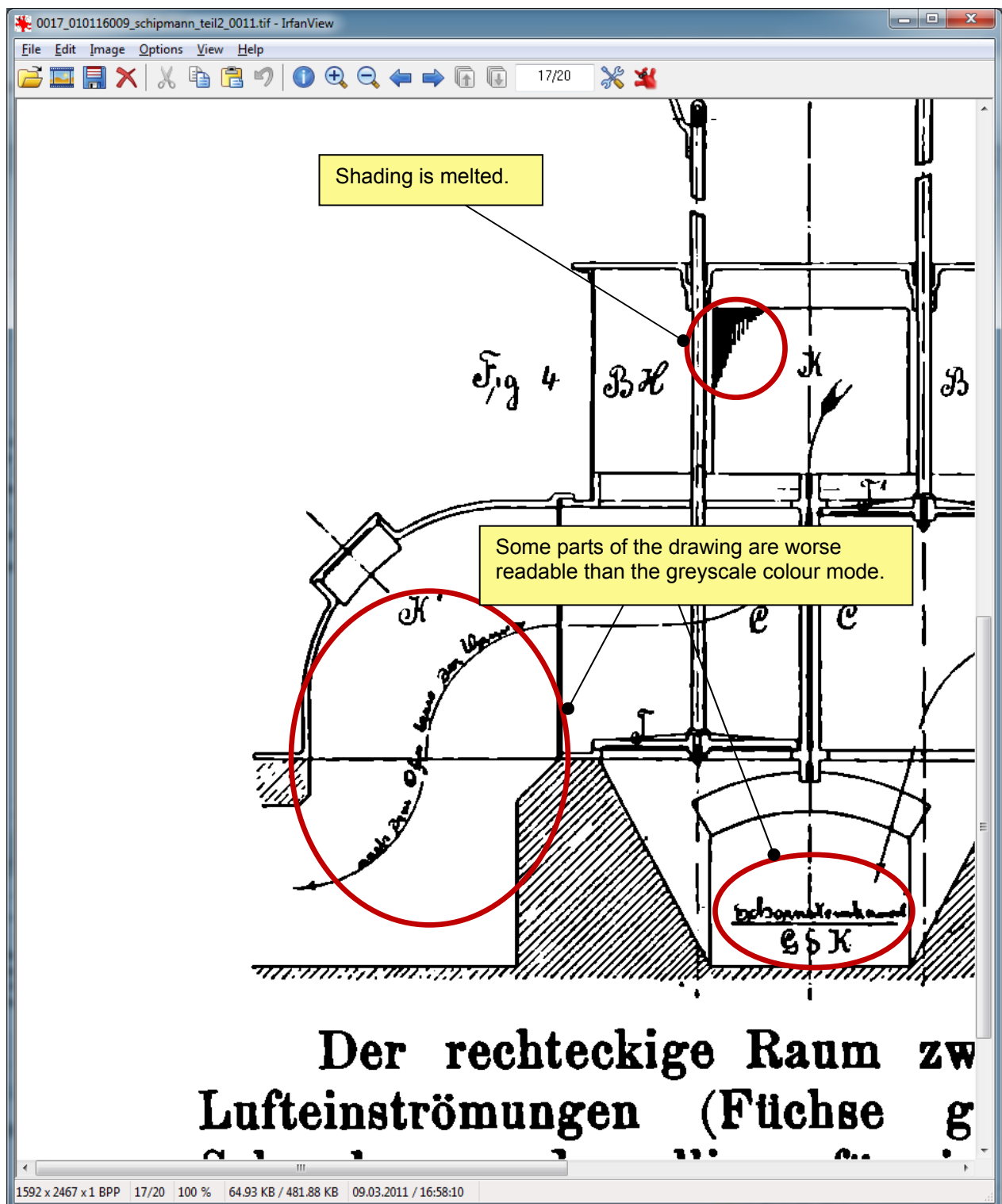


Figure B10 - ScanTailor processed image with colour mode settings: *Black and White*, Thinner/Thicker slider = 0
 Advantages: clean background, smaller file size,
 Disadvantages: types are smoothed with a little more unpleasant readability of the types and the handwriting in the drawing, bad shading resolution in the drawing

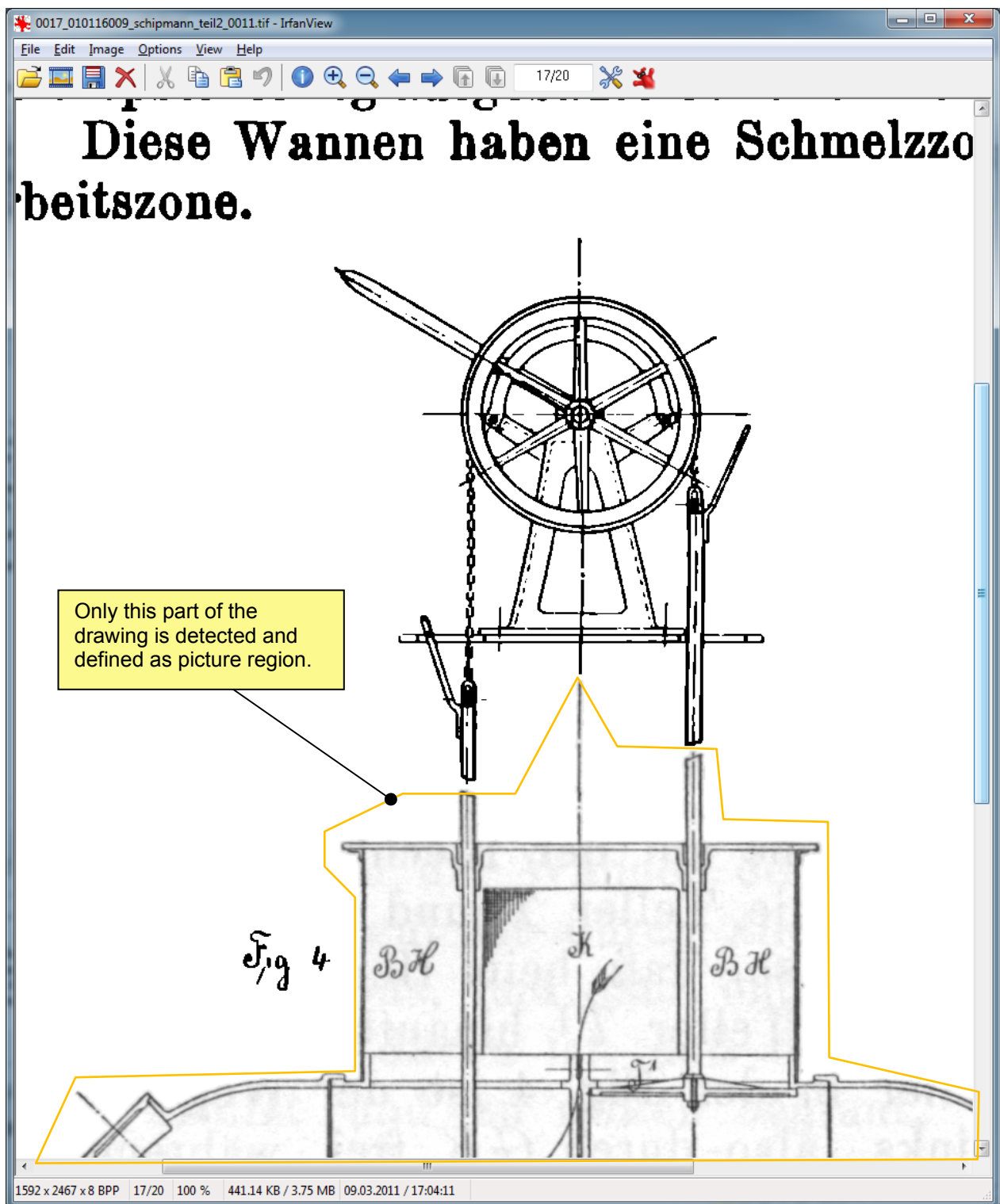


Figure B11 - ScanTailor processed image with colour mode settings: *Mixed*, Thinner/Thicker slider = 0. The upper parts of the drawing and the text are black and white (interpreted as text region) the lower part is in greyscale (interpreted as picture region), resizing the picture region is necessary by using the *Picture Zones* tab

B4 Influence of the colour mode settings on the quality of photographs

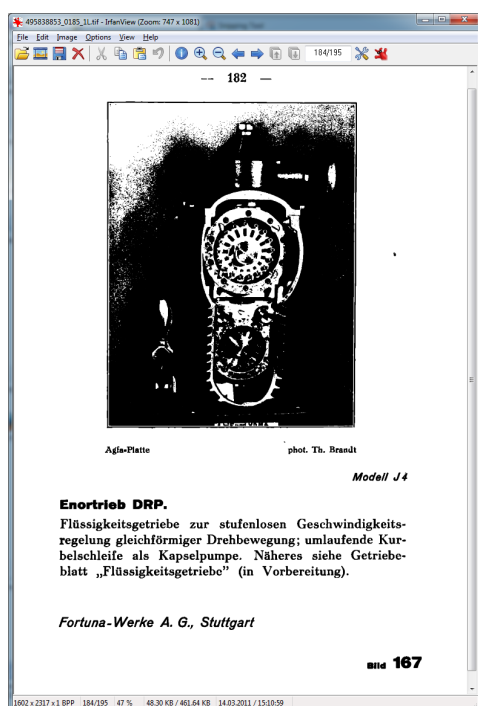


Figure B12 - ScanTailor colour mode setting: **Black and White**
Quality of the photograph is very bad,
Output file size of TIFF-ZIP is 48KB.
Do not use this setting for such types of
sources!



Figure B13 - ScanTailor colour mode setting: **Grayscale**
Rear side is shining through
Output file size of TIFF-ZIP is
2,41MB.



Figure B14 - ScanTailor colour mode setting: **Mixed mode**
Quality of the photograph is much better.
Output file size of TIFF-ZIP is 1,24MB.

Annex C – Examples for working with *ABBYY FineReader*

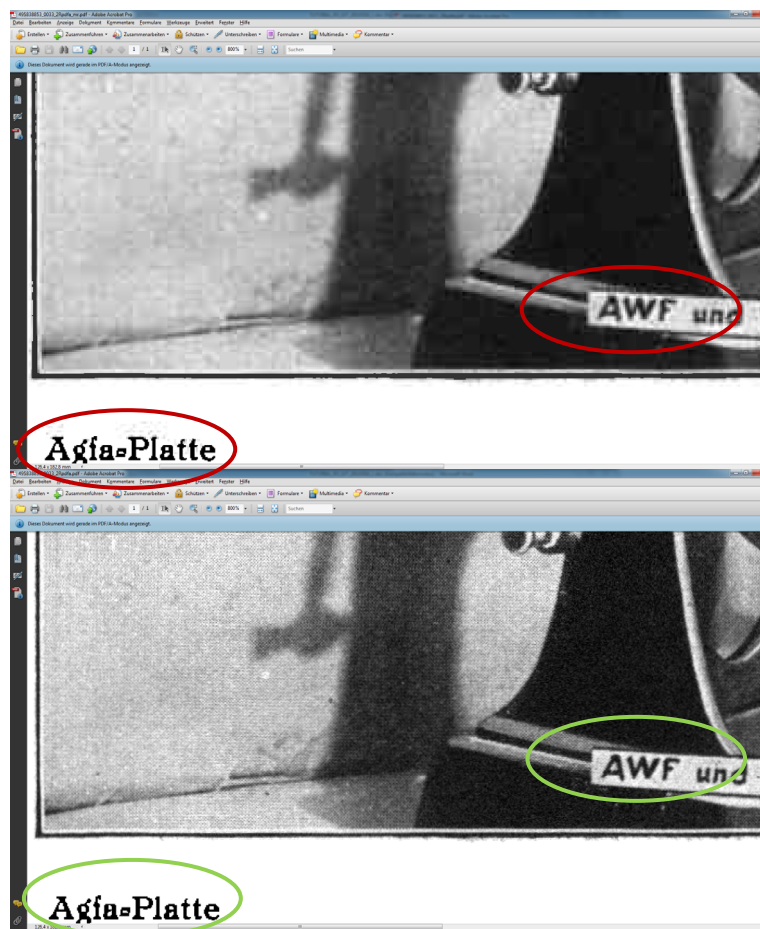


Figure C1 – *ABBYY FineReader*'s PDF/A export option *Mixed Raster Content* switch on (upper figure) and switch off (lower figure).
Do not use this option, because it causes a JPEG compression!