IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

IMPACT Storage Estimator: Tutorial (for Version 4.0, February 2011)

Fedor Bochow, Munich DigitiZation Center (MDZ)

Table of Contents

Introduction

This is a short guide on how to use the offline version of the IMPACT Storage Estimator (ISE) that is part of the IMPACT Decision Support Tools. It complements the implied documentation of the ISE itself.

In addition to this offline version of the ISE, there is also a simplified online version which will be made available shortly through the IMPACT Website.

The ISE is provided for the IMPACT Project by the Munich DigitiZation Center (MDZ) at the Bavarian State Library. Of course, this estimator is not the work of a single person or institution. Rather, it is based on the experience and input of several colleagues from many IMPACT institutions. Despite this, we would like to especially thank Ed I Bremner from the University of Bath who provided very useful input to improve the ISE.

Although care was taken in producing this estimator, please use it at your own risk. IMPACT does not guarantee that it free from errors or omissions.

We would very welcome all kinds of feedback, especially related to the

- Design and user guidance
- Completeness of information
- Correctness of calculation and content

You can contact the authors through the IMPACT Helpdesk:

http://www.impact-project.eu/helpdesk

What is the IMPACT Storage Estimator (ISE)?

The IMPACT Storage Estimator enables users and decision makers to establish the storage requirements, or in other words, to estimate the amount of storage and storage media needed for master files, derivative files and OCR files during their (mass) digitisation projects. It is optimised for images containing text.

The offline version of the IMPACT Storage Estimator is a spreadsheet that can be easily filled in by the end-user. The tool was created with Microsoft Excel 2007 (compatibility mode) and should operate with other Excel versions as well. Besides, it should – with limitations – also be usable with the open-source application OpenOffice.org. The Estimator can be freely accessed via the IMPACT website. The ISE contains a short introduction, the estimator itself and the implied documentation mainly consisting of guiding tool tips.

The IMPACT Storage Estimator was previously called IMPACT Storage Calculator, but renamed because the calculation involves several indicative values in the calculation, making it impossible to produce exact results. One example of this is the JPG compression rate, which cannot be easily standardised and therefore differs depending on the image processing software that is used. We must therefore stress that although the tool is based on a sound foundation, it can only produce an estimation of the storage needs!

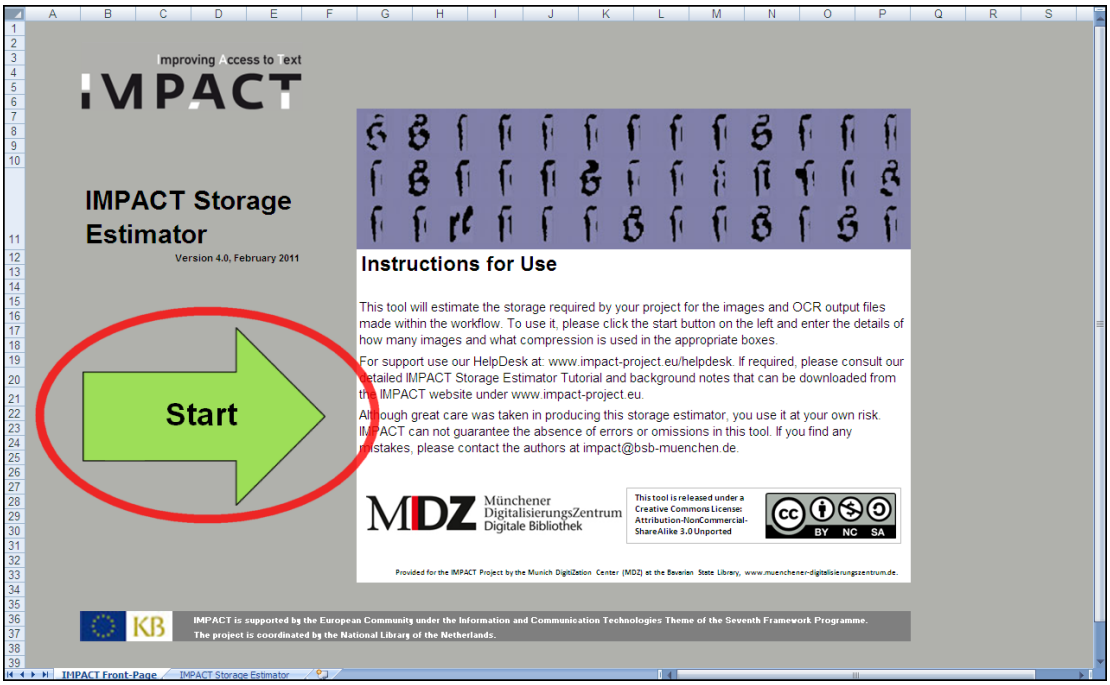
Who is the intended audience for the ISE?

Of course, the estimator can be downloaded and used by everyone via the internet. However, it is mainly designed for decision makers and staff involved in the (mass) digitisation of cultural heritage material. It will be of particular interest to institutions like museums, libraries and archives and also for those who are new to, or inexperienced in (mass) digitisation.

How to use the ISE?

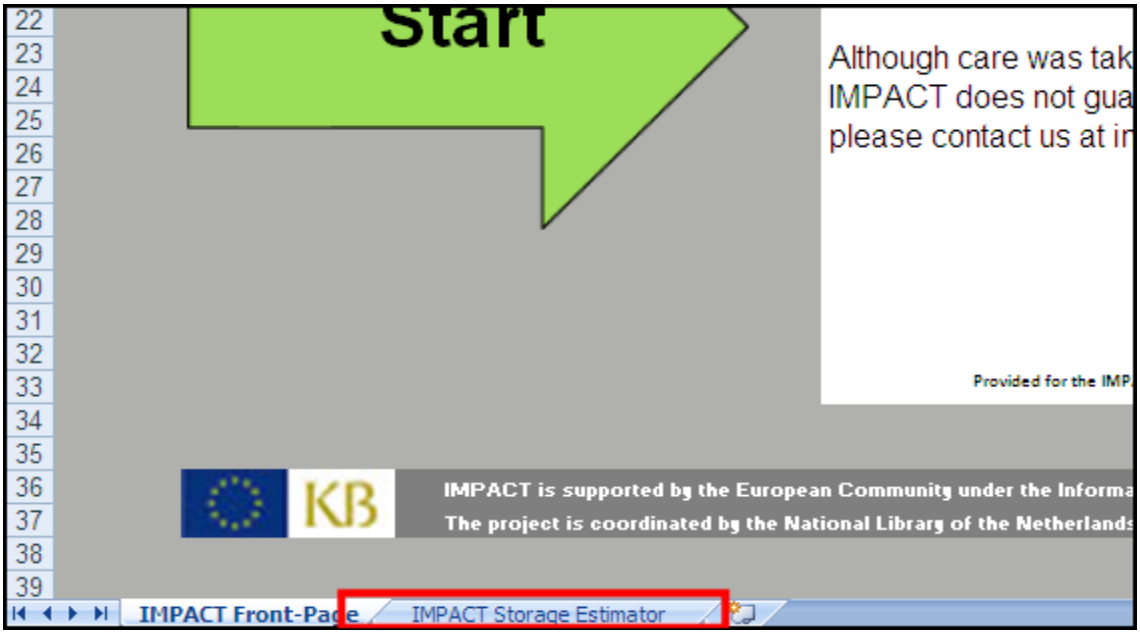
Front-page

When you open the Excel file you will see a welcome page with a short introduction. To start a calculation, the user has to click on the green start button and then enter the details of the number of images and the compression used in the appropriate boxes.

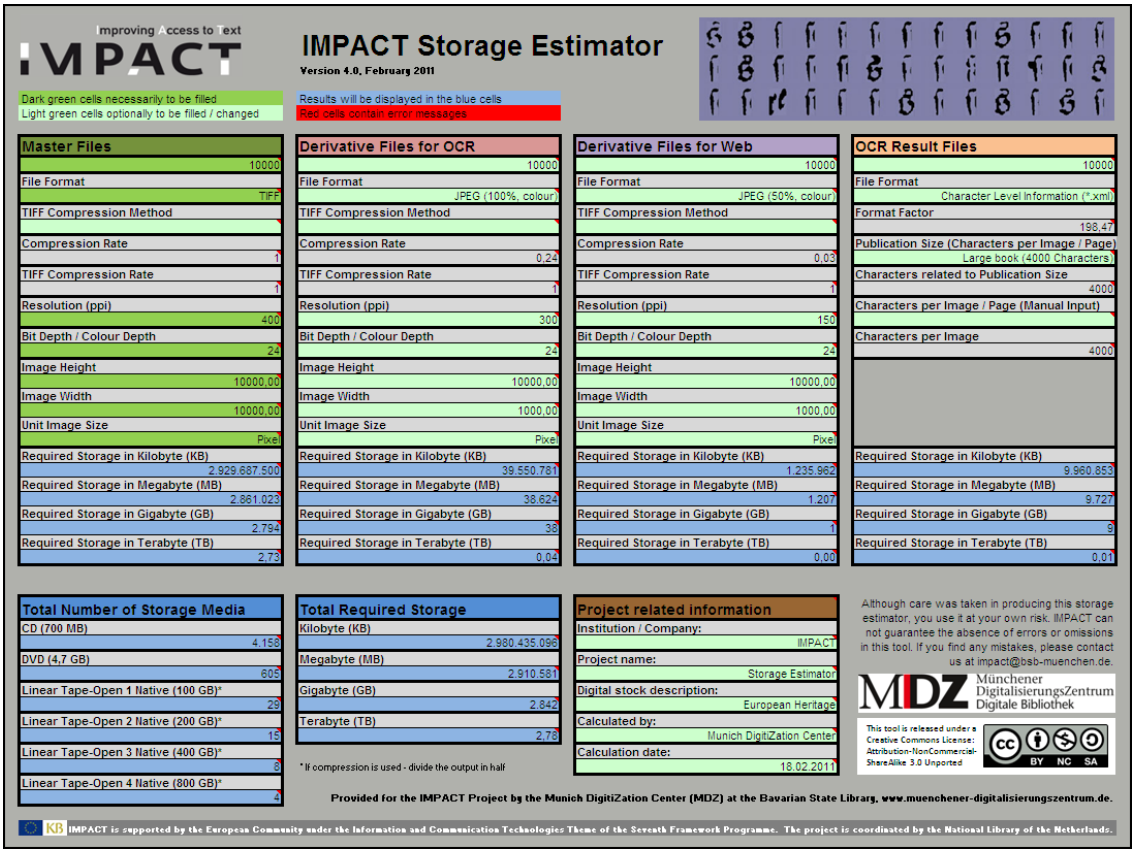


Welcome page of the IMPACT Storage Estimator (ISE). Just click on the green arrow to start a calculation.

Alternatively, the user can click on the worksheet tab that is located below on the left side and named “IMPACT Storage Estimator” to open the calculation page:



Calculation Sheet



IMPACT Storage Estimator

Version 4.0, February 2011

Dark green cells necessarily to be filled
Light green cells optionally to be filled / changed

Results will be displayed in the blue cells
Red cells contain error messages

Master Files		Derivative Files for OCR		Derivative Files for Web		OCR Result Files	
	10000		10000		10000		10000
File Format	TIFF	File Format	JPEG (100%, colour)	File Format	JPEG (50%, colour)	File Format	Character Level Information (*.xml)
TIFF Compression Method		TIFF Compression Method		TIFF Compression Method		Format Factor	198,47
Compression Rate	1	Compression Rate	0,24	Compression Rate	0,03	Publication Size (Characters per Image / Page)	Large book (4000 Characters)
TIFF Compression Rate	1	TIFF Compression Rate	1	TIFF Compression Rate	1	Characters related to Publication Size	4000
Resolution (ppi)	400	Resolution (ppi)	300	Resolution (ppi)	150	Characters per Image / Page (Manual Input)	
Bit Depth / Colour Depth	24	Bit Depth / Colour Depth	24	Bit Depth / Colour Depth	24	Characters per Image	4000
Image Height	10000,00	Image Height	10000,00	Image Height	10000,00		
Image Width	10000,00	Image Width	1000,00	Image Width	1000,00		
Unit Image Size	Pixel	Unit Image Size	Pixel	Unit Image Size	Pixel		
Required Storage in Kilobyte (KB)	2.929.687.500	Required Storage in Kilobyte (KB)	39.550.781	Required Storage in Kilobyte (KB)	1.235.962	Required Storage in Kilobyte (KB)	9.960.853
Required Storage in Megabyte (MB)	2.861.023	Required Storage in Megabyte (MB)	38.624	Required Storage in Megabyte (MB)	1.207	Required Storage in Megabyte (MB)	9.727
Required Storage in Gigabyte (GB)	2.794	Required Storage in Gigabyte (GB)	38	Required Storage in Gigabyte (GB)	1	Required Storage in Gigabyte (GB)	9
Required Storage in Terabyte (TB)	2,73	Required Storage in Terabyte (TB)	0,04	Required Storage in Terabyte (TB)	0,00	Required Storage in Terabyte (TB)	0,01

Total Number of Storage Media	
CD (700 MB)	4.158
DVD (4,7 GB)	605
Linear Tape-Open 1 Native (100 GB)*	29
Linear Tape-Open 2 Native (200 GB)*	15
Linear Tape-Open 3 Native (400 GB)*	8
Linear Tape-Open 4 Native (800 GB)*	4

Total Required Storage	
Kilobyte (KB)	2.980.435.096
Megabyte (MB)	2.910.581
Gigabyte (GB)	2.842
Terabyte (TB)	2,78

* If compression is used - divide the output in half

Project related information	
Institution / Company:	IMPACT
Project name:	Storage Estimator
Digital stock description:	European Heritage
Calculated by:	Munich DigitiZation Center
Calculation date:	18.02.2011

Although care was taken in producing this storage estimator, you use it at your own risk. IMPACT can not guarantee the absence of errors or omissions in this tool. If you find any mistakes, please contact us at impact@bsb-muenchen.de.

MDZ Münchener DigitalisierungsZentrum Digitale Bibliothek

This tool is released under a Creative Commons License: Attribution-NonCommercial-ShareAlike 3.0 Unported

Provided for the IMPACT Project by the Munich DigitiZation Center (MDZ) at the Bavarian State Library, www.muenchener-digitalisierungszentrum.de.

IMPACT is supported by the European Community under the Information and Communication Technologies Theme of the Seventh Framework Programme. The project is coordinated by the National Library of the Netherlands.

The IMPACT Storage Estimator calculation sheet. Data cells in green require input by the user; data cells in blue contain calculation results.

The calculation sheet consists of several areas with three types of cells:

- user controlled input fields (green)
- automatically filled in output fields (blue)
- unchangeable fields with static information (grey)

The dark green fields in the Master Files area have to be filled in. The light green fields – e.g. for “OCR Result Files” and “Project related information” – are optional.

Based on the input in the green fields, all calculation results will be provided in blue cells. Consolidated results will be displayed in the areas “Total Number of Storage Media” and “Total Required Storage”.

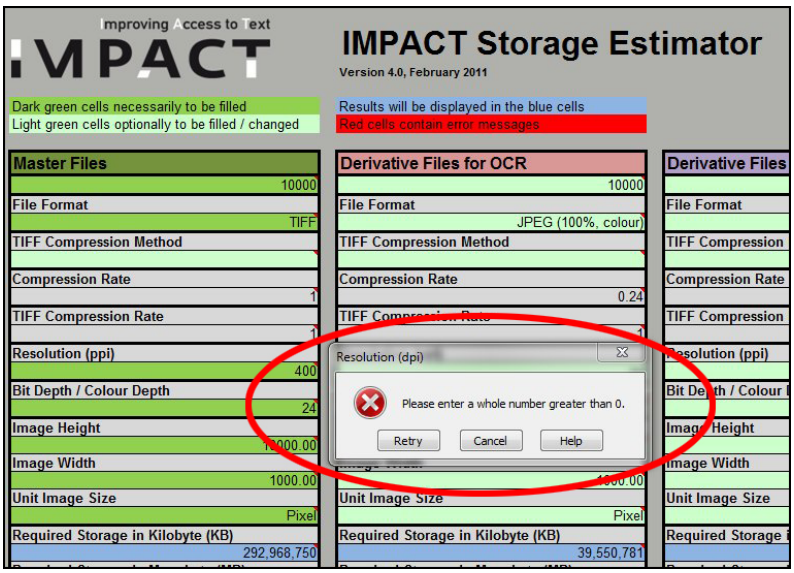
Hint: Apart from the input fields, all cells are protected by default to avoid unwanted destruction of formula. This was made to guarantee proper functioning of the ISE. If required, use Excel’s “Unprotect Sheet” function to make the content of all cells available.

User Assistance

There is a large variety of user-assistance during the calculation process consisting of guiding tool tips, windows with alarm messages and red cells containing error descriptions.

Alarm messages

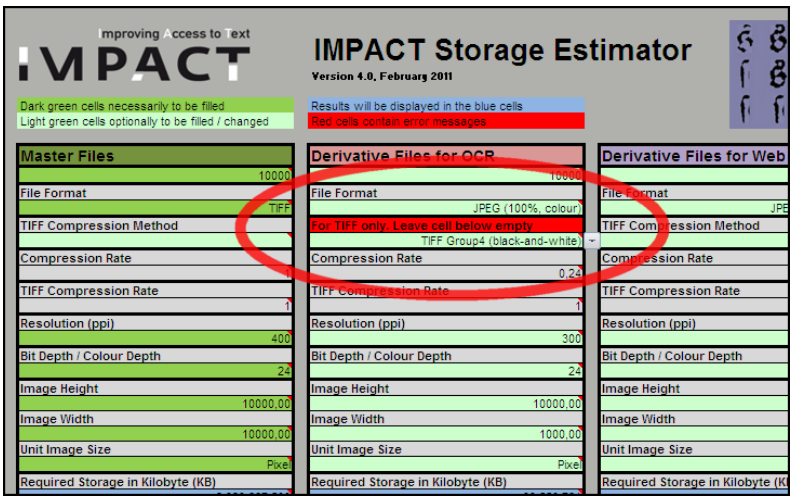
Alarm messages appear if the user tries to insert incorrect content in a green input field. For example, this window appears, if a user inserts text in a field that requires a numerical value:



If such a message appears, the user has to click on the Cancel button and correct the data entry.

Red cells

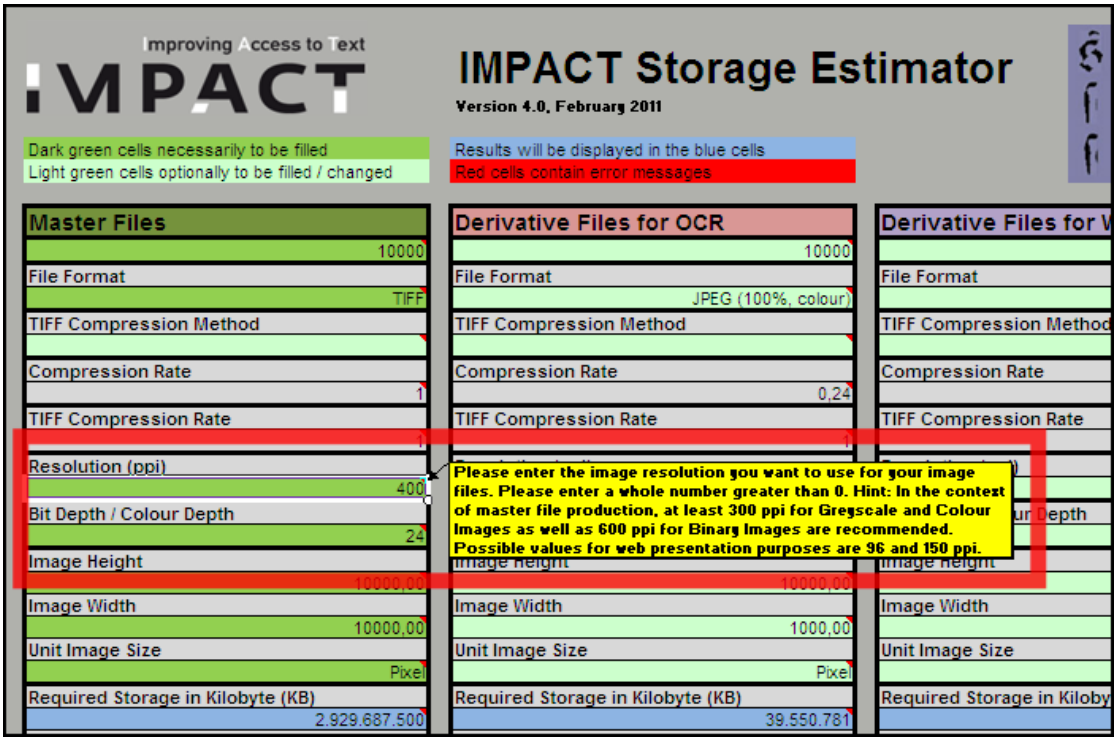
If you enter an incorrect combination of values, the cells will change colour to red and display error messages asking the user to correct the input. In the following example, someone accidentally tried to combine TIFF compression with the JPEG file format:



If such a message appears, the user must go back and re-enter the data correctly. Of course, error messages will disappear after correction.

Fly-out tool tips

Every input and output field has a fly-out tool tip. These tool tips explain to the user the kind of input that is needed or what output will be displayed. The user has just to move the mouse over the according field and a fly-out comment in yellow appears. Messages disappear again when the user moves the mouse away.



Just move the mouse over the input / output cells to get guiding tool tips (yellow).

Master Files Area

In this area, the user has to define the number and type of master files. Digital master files are mostly images directly created from the source material and normally also used for long-term preservation.

Of course, all master files have to be of similar nature to get a reliable result in the calculation. If there are different master file-types within a digitisation project; e.g. bitonal, greyscale and colour images as well, it is recommended to make different calculations for every file-type and to manually add the results.

The user has to insert the following information:

- ***Number of Master Files***

  Enter the estimated number of master files for the digitisation project here. Normally, this value will be closely connected or identical to the page number of the printed source material.
- ***File Format***

Enter the master file format here. Please note: RGB Baseline TIFF v6 is the most common format for master files. If JPEG files are used as master files, then they should only be used at the highest quality possible, as any loss of quality (compression) will be exaggerated when copies / derivatives are made. However, the following common file formats are supported by the ISE:

- TIFF
- GIF (black-and-white)
- GIF (colour; lossless up to 256 colours)
- GIF (greyscale)
- JPEG (100%, colour)
- JPEG (100%, greyscale)
- JPEG (50%, colour)
- JPEG (50%, greyscale)
- JPEG2000 (lossless, colour and greyscale)
- PNG (black-and-white, best compression)
- PNG (colour, best compression)
- PNG (greyscale, best compression)
- BMP (uncompressed)

The percentage value in brackets with the JPEG value represents the nominal quality (based on values created by two different image-processing tools; for more details, please read the tools below).

Master Files
10000
File Format
TIFF
TIFF Compression Method
Compression Rate
1
TIFF Compression Rate
1
Resolution (ppi)
400
Bit Depth / Colour Depth
24
Image Height
10000,00
Image Width
10000,00
Unit Image Size
Pixel

- ***Resolution (ppi)***

  Enter the image resolution for the image files here. *Hint: In the context of master file production, at least 300 ppi for Greyscale and Colour Images as well as 600 ppi for Binary Images are recommended. Possible values for web-presentation purposes are 96 and 150 ppi.*

- ***Bit Depth / Colour Depth***

  Enter the bit depth (= colour depth) for the image files here. Some common examples:

  - "1" for Binary Images
  - "8" for 256 tone Greyscale or colour Images
  - "24" for RGB (8bit per channel) True Colour Images and
  - "32"-"48" for extended bit depth images.

  Attention: It is possible to enter values that don't make sense, e.g. 24bit or more in combination with GIF files, which is not possible.

- ***Image Size (Height, Width) & according measure unit***

  Enter the height and width of the image files and the appropriate measure unit here. Some examples (Image Height x Image Width & Unit Image Size):

  - Choose 15 x 12 inches for Folio Format - e.g. a newspaper
  - 12 x 9.5 inches for Quarto Format - e.g. a large book
  - 9 x 6 inches for Octavo Format - e.g. an average book
  - 7.36 x 5 inches for Duodecimo Format - e.g. a small book

  The following common linear measurements are supported: pixel, inches, cm, mm. The default measure unit are pixels.

The following information may be optionally entered by the user:

- ***TIFF Compression Method***

  The TIFF compression method may only be chosen if the TIFF file format is selected beforehand. It has to be left blank if no compression is used. Only the following common TIFF compression methods are valid in the context of the ISE:

  - TIFF Group4 (for black-and-white images)
  - TIFF LZW (for black-and-white images)
  - TIFF LZW (for colour and greyscale images)
  - TIFF ZIP (for black-and-white images)
  - TIFF ZIP (for colour images)
  - TIFF ZIP (for greyscale images)

The Fields ***Compression Rate*** and ***TIFF Compression Rate*** are automatically filled in. The underlying values are calculated on the basis of some typical example files, which were created with different equipment from a range of source material (books, newspapers) published from the 18th century onwards.

To undertake this image compression, two tools were used:

- IrfanView (http://www.irfanview.net)
- XnView (http://www.xnview.com)

In most cases, both tools produced identical or nearly identical results (file sizes were very close in size after converting from one file format to another). In some cases, there were slightly larger differences. To establish a reliable value for compression, the average size of the files produced by the tools was chosen. If required, test material can be provided for these calculations. More background information on this can be found when activating the “Background Compression & Format” tab of the ISE, which is faded out and protected by default. To unprotect the cells please consult the hint in the section “Calculation Sheet”.

Derivative Files for OCR Area

In this area, the user is able to define the number and type of derivative files for OCR. Please note: In case there are no derivative files in the project, the value in the according field has to be “0” to avoid incorrect results. Normally, derivative files are directly created from the master files.

Of course, all derivative files have to also be of similar nature to get a reliable estimation at the end. If there are different file-types within a digitisation project – e.g. bitonal, greyscale and colour images as well, it is recommended to run the estimator for every file-type and to then manually add the results.

Derivative Files for OCR
10000
File Format
JPEG (100%, colour)
TIFF Compression Method
Compression Rate
0,24
TIFF Compression Rate
1
Resolution (ppi)
300
Bit Depth / Colour Depth
24
Image Height
10000,00
Image Width
1000,00
Unit Image Size
Pixel

These fields may be optionally filled in by the user:

- ***Number of Derivative Files for OCR***
  Enter the estimated number of derivative files for OCR here. Typically, this value will be identical to the number of master files.
- ***File Format***
  Enter the derivative file format here. Details of the available file formats can be found in the section “Master Files Area”. However, it is worth mentioning, that the common master file format TIFF does not necessarily lead to better OCR quality. Quite the contrary, in some cases compressed JPEG files (down to 25 percent!) are generating better or equal OCR results in comparison to the TIFF master files. Unfortunately, we do not have detailed scientific evaluation in this context to provide reliable values here.
- ***TIFF Compression Method***
  Details can be found in the section “Master Files Area”.
- ***Resolution (ppi)***
  Details can be found in the section “Master Files Area”.

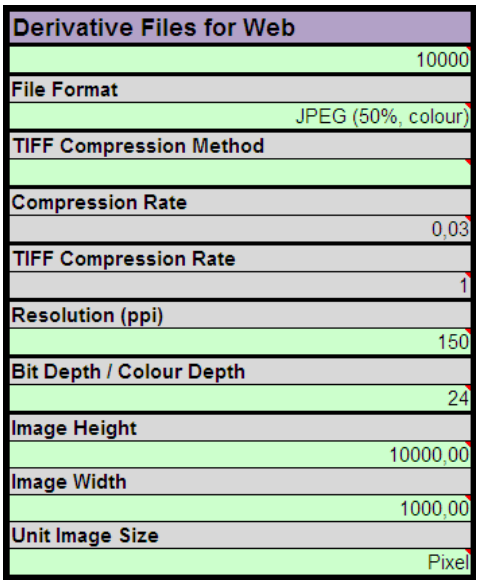
- ***Bit Depth / Colour Depth***
  Details can be found in the section “Master Files Area”.
- ***Image Size (Height, Width) & according measure unit***
  Details can be found in the section “Master Files Area”.

Details related to the automatically filled in fields ***Compression Rate*** and ***TIFF Compression Rate*** can be found in the section “Master Files Area”.

Derivative Files for Web presentation Area

In this area, the user is able to define the number and type of derivative files for web presentation. Please note: If there are no derivative files used in the project, the value in this field has to be “0” to avoid an incorrect estimation. Normally, derivative files are directly created from the master files.

Of course, all derivative files have to also be of similar nature to get a reliable estimation at the end. If there are different file-types within a digitisation project – e.g. bitonal, greyscale and colour images as well, it is recommended to run the estimator for every file-type and to then manually add the results.



These are the following fields to be optionally filled in by the user:

- ***Number of Derivative Files for OCR***
  Enter the estimated number of derivative files for web presentation here. Typically, this value will be identical to the number of master files.
- ***File Format***
  Enter the derivative file format here. Details of the available file formats can be found in the section “Master Files Area”. However, it is worth mentioning here that derivative files for web presentation are normally compressed files to cut loading times and save on resources.

- ***TIFF Compression Method***
  Details can be found in the section “Master Files Area”.
- ***Resolution (ppi)***
  Details can be found in the section “Master Files Area”.
- ***Bit Depth / Colour Depth***
  Details can be found in the section “Master Files Area”.
- ***Image Size (Width, Height) & according measure unit***
  Details can be found in the section “Master Files Area”.

Details related to the automatically filled in fields ***Compression Rate*** and ***TIFF Compression Rate*** can be found in the section “Master Files Area”.

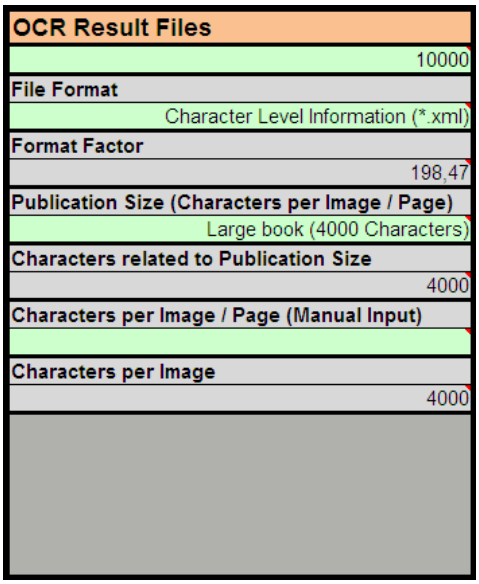
OCR Result Files Area

In this area the user is able to define the number and type of OCR Result Files (i.e. files containing the recognised text created from the scanned / photographed images). Please note: If you don’t have any of these files in the project, the value in this field has to be “0” to avoid an inaccurate estimation.

These fields may be optionally filled in by the user:

- ***Number of OCR Result Files***
  Enter the estimated number of OCR result files for the digitisation project here. Typically, the value will be identical to the number of image files entered under “Area related to Derivative Files for OCR”.
- ***File Format***
  Select the file format you plan to use for the OCR result files (plain text or text with information on word / character level) here. Only the following file formats are valid in the context of the ISE:
  - Unformatted Text (*.txt)
  - Word Level Information (*.xml)
  - Character Level Information (*.xml)
- ***Publication Size (Characters per Image / Page)***
  Select the publication size related to characters per image / page here. If you use the field below - “Characters per Image / Page”, then you must leave this field blank. Only the following entries are valid for use in the ISE:
  - Small book (with 2.000 characters per page)
  - Average book (with 3.000 characters per page)
  - Large book (with 4.000 characters per page)
  - Small newspaper (with 8.000 characters per page)
  - Average newspaper (with 10.000 characters per page)
  - Large newspaper (with 12.000 characters per page)
- ***Characters per Image / Page (Manual Input)***

If available, the (estimated) number of characters per image / page can be manually entered here. Otherwise, the cell should be left blank and the field above - “Publication Size (Characters per Image / Page)” - should be used instead.



The fields;

- ***Format Factor***
- ***Characters related to Publication Size***
- ***Characters per Image***

are automatically filled in.

The underlying values for the field ***Format Factor*** are calculated on the basis of some typical example files created with different equipment from different source material (books, newspapers) published from the 19th century onwards. The tool used for creating the different OCR result files (plain text or text with information on word / character level) was ABBYY FineReader 7.0. Test results can be provided. Background information can be found by activating the “Background OCR & Format” tab of the ISE, which is faded out and protected by default. To unprotect the cells please consult the hint in the section “Calculation Sheet”.

Result areas

Based on the given input, the user will find all relevant output information in the blue cells. Values are provided for Master Files, Derivative Files and OCR Result Files.

The most relevant results will normally be the consolidated values in the following areas:

- ***Total Required Storage***

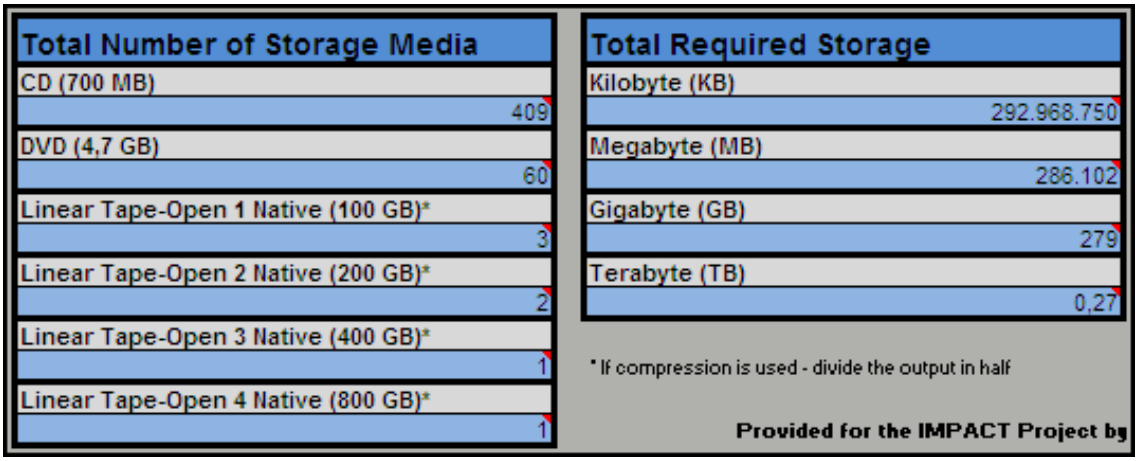
  The storage requirements for the project are displayed here, in the units: Kilobyte (KB), Megabyte (MB), Gigabyte (GB) and Terabyte (TB).

- ***Total Number of Storage Media***

  The required number of storage media for the project is displayed here. The following common storage media types are supported:

  - CD (700 MB capacity)
  - DVD (4,7 GB capacity)
  - Linear Tape-Open 1 Native (100 GB capacity)
  - Linear Tape-Open 2 Native (200 GB capacity)
  - Linear Tape-Open 3 Native (400 GB capacity)
  - Linear Tape-Open 4 Native (800 GB capacity)

Hint for Linear Tape-Open: Please divide the output in half, if compression is used.



Total Number of Storage Media	
CD (700 MB)	409
DVD (4,7 GB)	60
Linear Tape-Open 1 Native (100 GB)*	3
Linear Tape-Open 2 Native (200 GB)*	2
Linear Tape-Open 3 Native (400 GB)*	1
Linear Tape-Open 4 Native (800 GB)*	1

Total Required Storage	
Kilobyte (KB)	292.968.750
Megabyte (MB)	286.102
Gigabyte (GB)	279
Terabyte (TB)	0,27

* If compression is used - divide the output in half

Provided for the IMPACT Project by

Area for Project related information

In this area, the user is able to add project related information for documentation purposes. You can add here the following details:

- Specification of the Institution / Company
- Name of the project
- Description of the digital stock
- Name of the calculator
- Calculation date (automatically filled in; manual input is also possible)

Project related information
Institution / Company:
IMPACT
Project name:
Storage Estimator
Digital stock description:
European Heritage
Calculated by:
Munich DigitiZation Center
Calculation date:
19.10.2010

Export of results

It is possible to export the estimation results or forward the estimation data in the following ways:

- The user can print out the Calculation Sheet, which has been optimised for an A4 paper size. Best results will be achieved when printing out the sheet in colour, but black-and-white printouts are also possible.
- The user can also create a PDF – or better PDF/A – file from the Calculation Sheet that is optimised for A4 paper. PDF is an open standard for optimised document exchange – independent of the application software, hardware, and operating system. PDF/A is optimal for long-term storage of the calculation sheet (e.g. for project documentation purposes). If PDF export is not supported by the version of Excel being used, one can use a free tool such as PDFCreator (http://www.pdfforge.org) to get similar results. *Hint: Once exported to PDF, you will not be able to change the calculation. Therefore, it is strongly recommended to store the Excel file as well.*

About this Document

Author – Fedor Bochow, Munich Digitization Center (MDZ) at the Bavarian State Library

Fedor Bochow joined the Munich DigitiZation Center at the Bavarian State Library in 2006, after working for a range of media companies and publishing houses. Since joining the MDZ, he has worked for several EU Projects on behalf of the Bavarian State Library. In 2008 he took over the management of the IMPACT team at the Bavarian State Library.

Revisions

Version	Status	Date Released	Lead Author
1.0	Pilot Release	22.10.2010	Fedor Bochow - Munich DigitiZation Center (MDZ) at the Bavarian State Library
2.0	Updated Version	18.02.2011	Fedor Bochow - Munich DigitiZation Center (MDZ) at the Bavarian State Library
4.0	First Release	31.03.2011	Fedor Bochow - Munich DigitiZation Center (MDZ) at the Bavarian State Library

About this release

This document is the tutorial that accompanies the offline version of the IMPACT Storage Estimator (ISE) that is part of the IMPACT Decision Support Tools. It completes the implied documentation of the ISE itself.

The IMPACT Storage Estimator has been released by the IMPACT project to assist practitioners and students in the (mass) digitisation of text and the use of Optical Character Recognition.

For your questions and feedback, please use our HelpDesk at: http://www.impact-project.eu/helpdesk.

In addition to this offline version of the ISE, there is also a simplified online version which will be made available shortly through the IMPACT Website.