READ
Recognition and Enrichment of Archival Documents

How To Export documents from Transkribus

Version v1.9.1
Last update of this guide: 28/11/2019

This guide explains how to export your documents from Transkribus. It will allow you to customise your export according to the file format and options you prefer.

Download the Transkribus Expert Client, or make sure you are using the latest version:

- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:

- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact:

- The Transkribus Team: email@transkribus.eu

Contents

The READ project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943.

Introduction

- If you want to work with your images and transcriptions outside of Transkribus, you can export your documents from the platform.
- Different export formats and features are available to suit your needs.

Export function

- To use the export function click on this symbol:

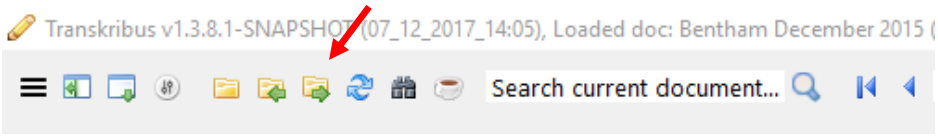

Figure 1 "Export document" button

- The following window will open up. You will see that there are two tabs offering Client export and Server export.
- If you choose the Server export option, the export will be processed on the Transkribus server and you will receive a link to download your files. The export will not slow your computer down and the process will not be interrupted if you switch your computer off.
- With Client export the files will be saved directly to your computer.

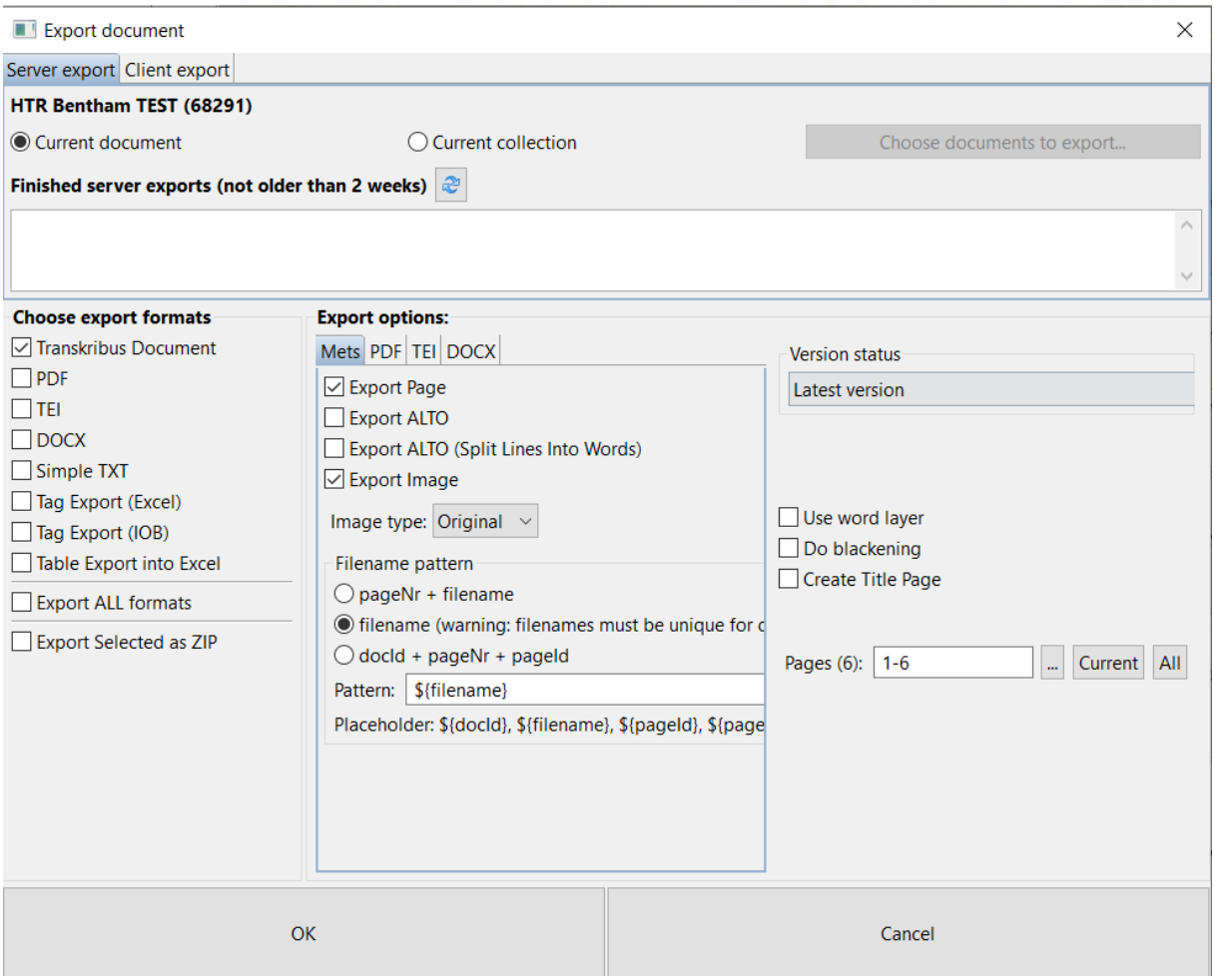

Figure 2 "Export document" window

Server Export

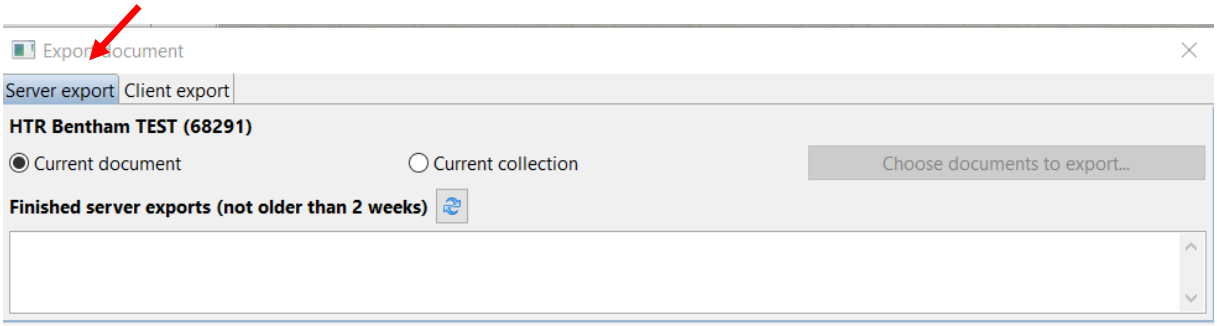

Figure 3 Server export

- Click the "Server" export tab, fill in your desired options and click OK.
- You can check the progress of your export by clicking the "Jobs" button in the "Server" tab.

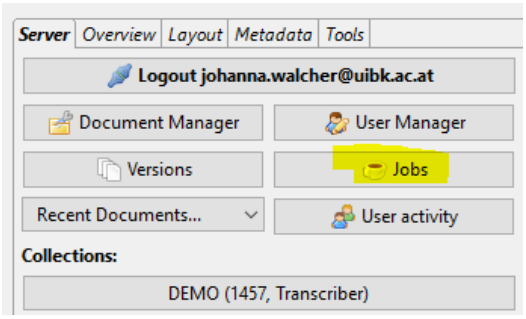

Figure 4 Click the "Jobs" button to check the progress of the export

- After the process has finished you will receive an email with the link to download the files. If you wish to pass the files onto someone else, you can forward the email to them.
- **Note:** The rest of this guide focuses on Client export. All of the options discussed below are also available in Server export.

Client export

- Click the "Client export" tab and fill in the below options:

Storage location

First of all, please choose where you would like to save the exported files. Type the file location in the "Base folder" box at the top of the "Export document" window.

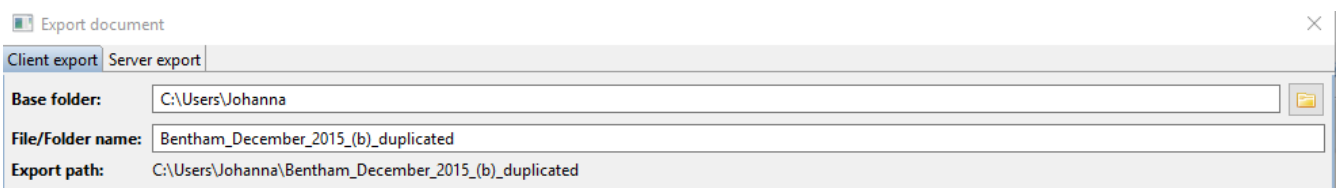

Figure 5 Indicate file location

Pages to be exported

- Select the number of pages you wish to export. You can export all the pages in your document, or just the current page.

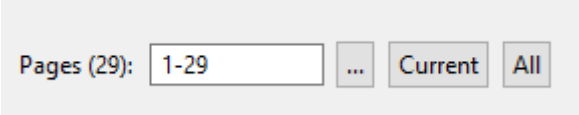

Figure 6 Choose the pages you would like to export

Export as Transkribus Document

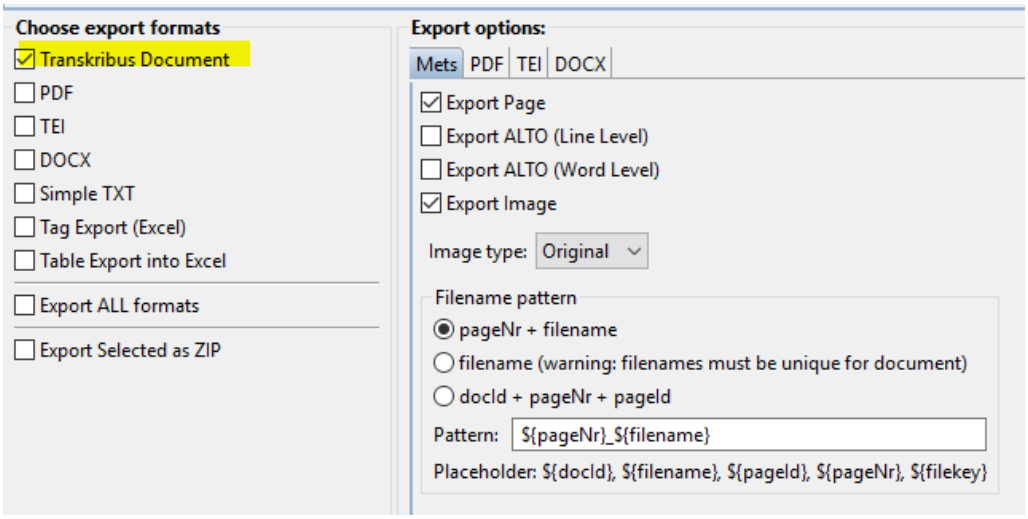

Figure 7 Export as Transkribus Document

- If you export your transcription as a Transkribus Document you will produce a METS (Metadata Encoding and Transmission Standard) file containing the links to PAGE, XMLs, ALTO and/or image files depending on which options you choose. A METS file is like a container which includes all the background information about a file. More detailed information about METS can be found at: http://www.loc.gov/standards/mets/

ALTO

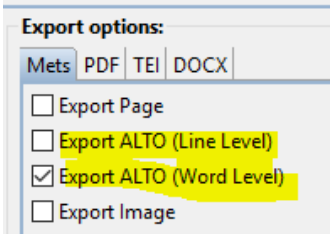

Figure 8 Export as ALTO document

- ALTO is a special output format which allows you to input the exported document into other programs working with this format. The format is similar to XML and works for OCR for example. It is often used in combination with METS for the description of the whole digitized object and creation of references across the ALTO files, e.g. description of the reading sequence. More information about ALTO can be found at: http://www.loc.gov/standards/alto/

- With the “Word Level” option Transkribus will divide the lines into words. The program does this by analysing the spaces between words, even if no word segmentation has been performed previously.

Filename pattern

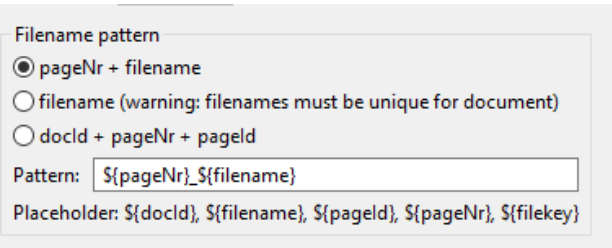

Figure 9 Filename pattern

- Under “Filename pattern” you can choose how the filename will be composed.
- **Note:** The second option “filename” is the standard one. With this option the exported file will have the same name as the document you imported. This is important if you want to match local transcripts with the images in Transkribus. So if you export a document, then adjust it externally and after that upload it to Transkribus again, the program will need to have two similar filenames in order to recognise the file properly.

Export as PDF

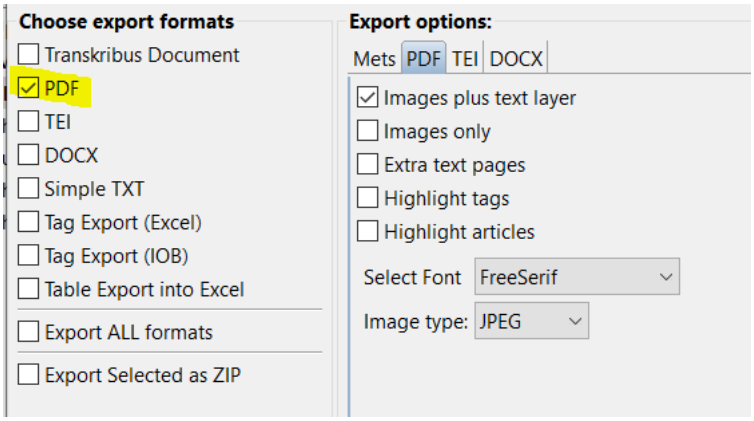

Figure 10 Export as PDF

- If you export a PDF file you can choose between “Images only” or “Images plus text layer”
  - **“Images only”**: you will produce a PDF file with the document as an image. This means that you will not see the transcribed text.

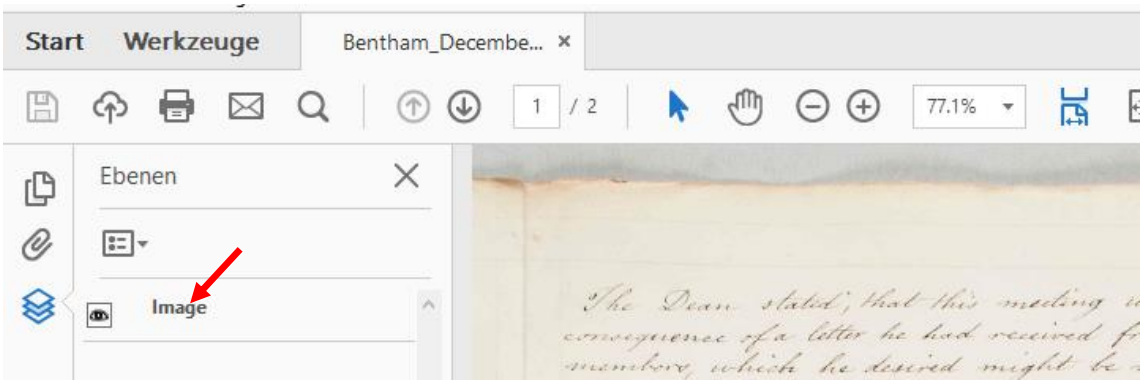

Figure 11 Exported PDF document

- **"Images plus text layer":** you will see two layers in the exported PDF document: OCR (the transcribed text) and image (image of the document).

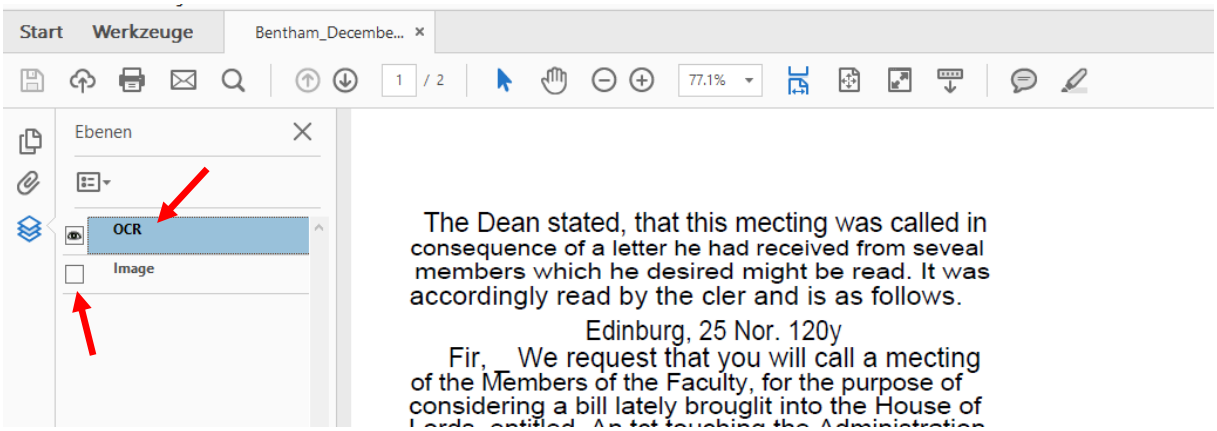

Figure 12 Layers in exported PDF

- **"Extra text pages":** if you scroll down in the exported PDF file, the transcribed text will appear after the image.
- **"Highlight tags":** this option will highlight any tags in your transcription. The tags will be shown in the same colours used in Transkribus. At the end of the document there will also be a symbol legend to explain the signification of the different colours.

Export as TEI

- This option is for people working with the **Text Encoding Initiative (TEI)**. Transkribus enables you to choose the **zones** you need.
- Furthermore you can choose between **line tags and line breaks**.
- The Text Encoding Initiative is a text-centric community of practice in the academic field of digital humanities, operating continuously since the 1980s. More information at: http://www.tei-c.org/index.xml

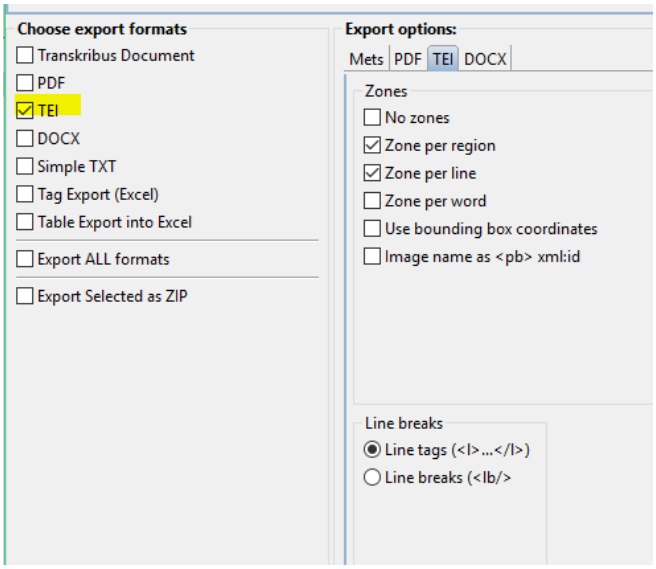

Figure 13 Export as TEI

Export as DOCX

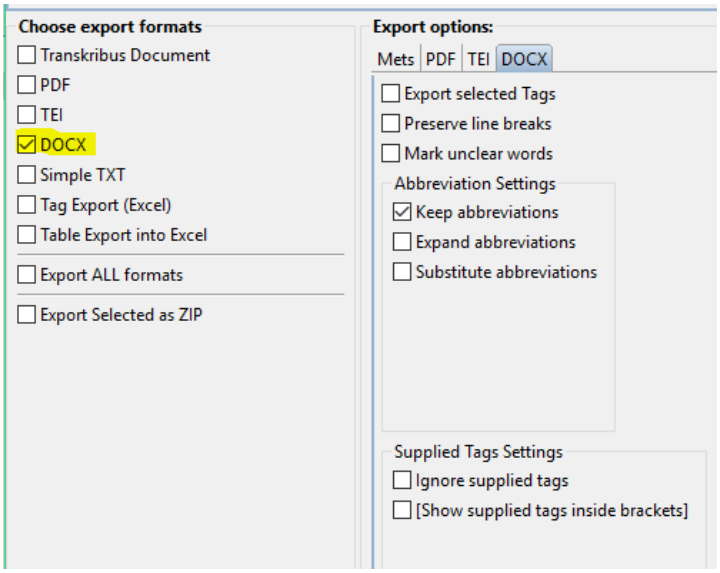

Figure 14 Export as DOCX

- By choosing this option you will get your transcriptions in Word files.
- You can select options relating to line breaks, abbreviations and more according to your needs.

Export as simple TXT

- If you do not usually work with Microsoft Word it is possible to export your transcription as a simple TXT file.

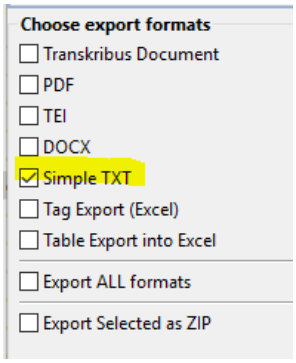

Figure 15 Export as TXT

Tag Export

If you would like to export the tags you assigned to your transcription, there are three possible options:

1. **Excel file:** select this option to produce an Excel file with individual tabs for each tag category and one tab with an overview of all the tags.

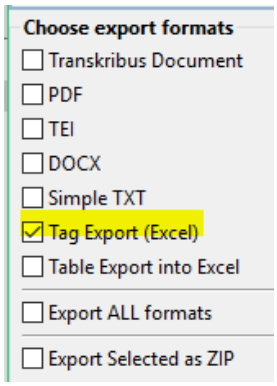

Figure 16 Tag export Excel

2. **PDF file:** select these options to highlight the tags in the exported PDF file.

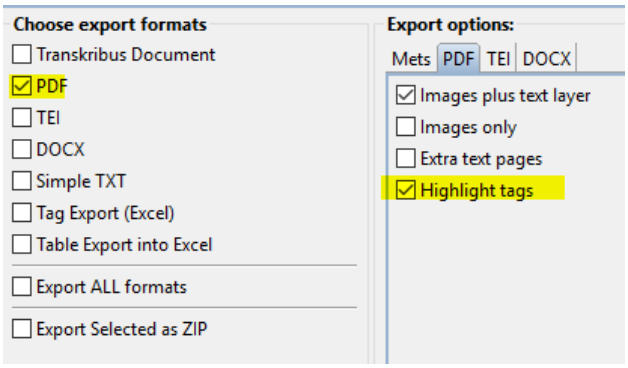

Figure 17 Tag export PDF

3. **DOCX file:** select these options to make the tags visible in the exported DOCX file.

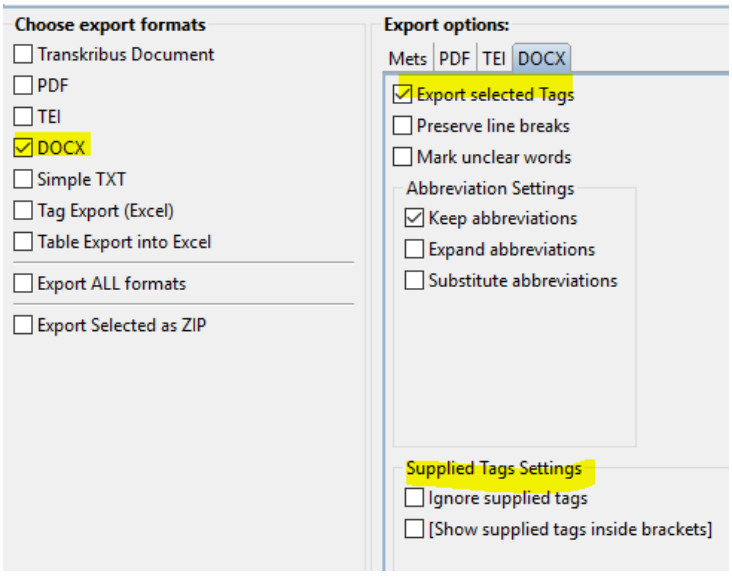

Figure 18 Tag Export DOCX

- After the export of the Word document please open it and do the following
  - Click on the paragraph button in the Home menu of Word.

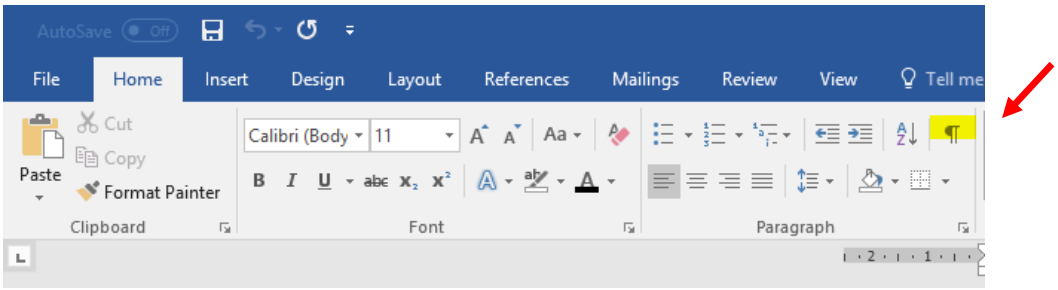

Figure 19 Paragraph button in Word

- Then go to “References” and choose “Insert Index”.

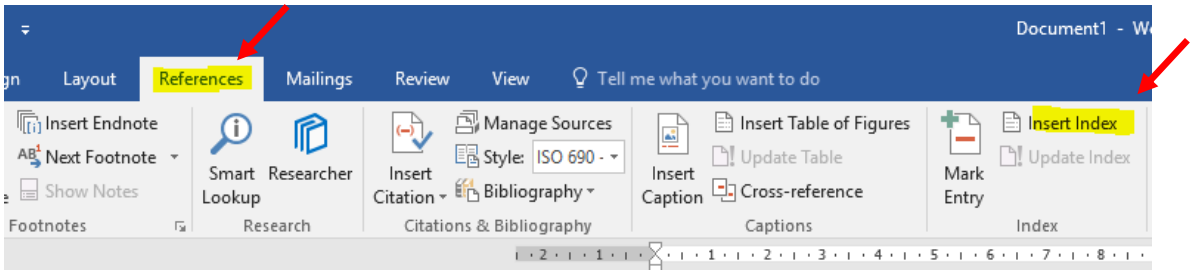

Figure20 Insert Index in Word

- The following Office window will open up:

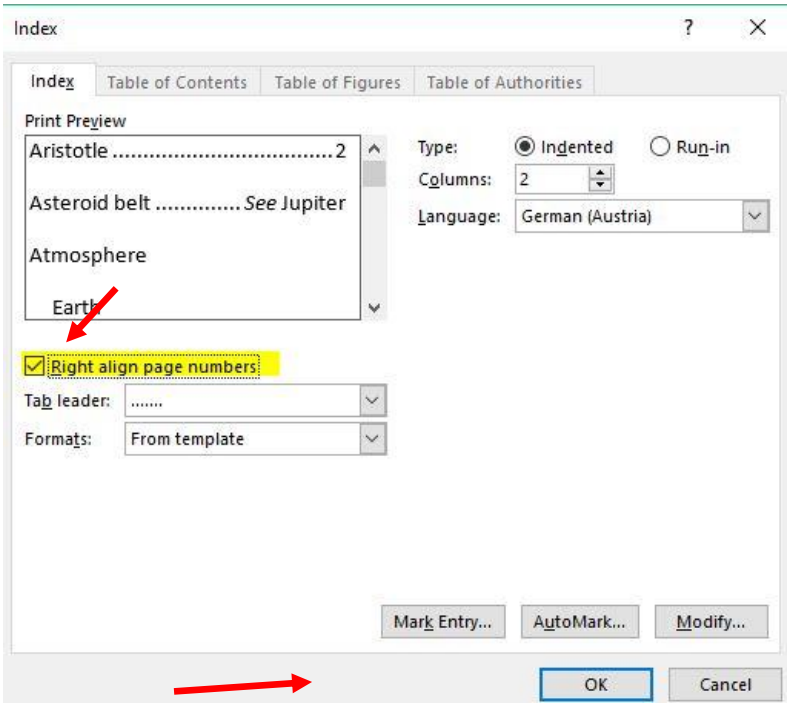

Figure 21 Index options in Word

- Select “Right align page numbers” and press “OK”
- A confirmation window will pop up, click “Yes”

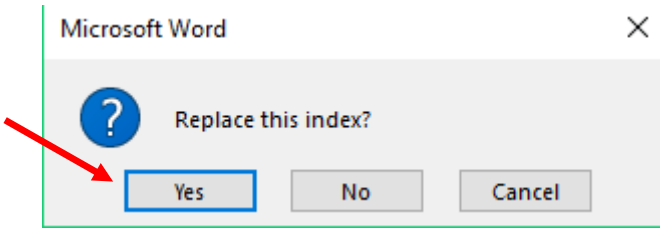

Figure 22 Confirming in Word

- An overview of the tags should now appear at the end of the document
- If the overview of the tags does not appear, click "Update Index". This should solve the problem.

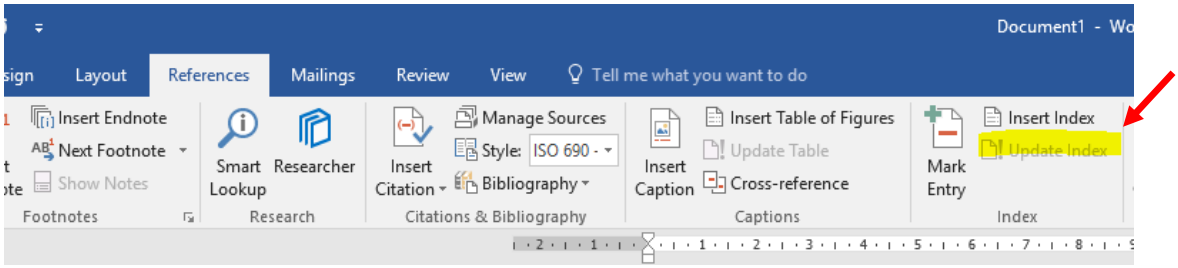

Figure 23 "Update Index" in Word

Table Export into Excel

For information concerning the export of tables please consult [How to Process Tables in Transkribus.]

More options

There are a few other export options to explore.

Export all formats

- This means that the transcription will be saved in the chosen location in all the available formats.
- If you choose "**Export ALL formats**" all the options will be ticked automatically

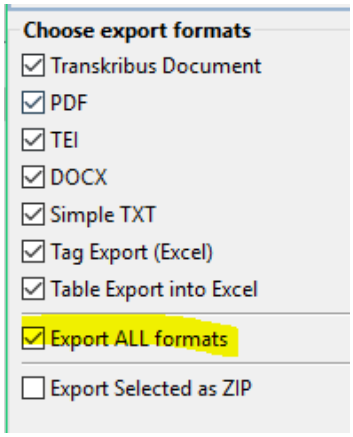

Figure 24 Export all formats

Use the version filter

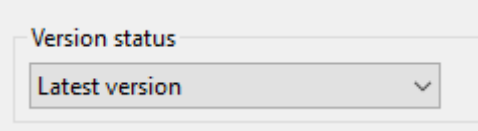

Figure 25 Choosing the version status

- This option makes it possible to export particular versions of the document.
- If you select “Ground Truth” for example, Transkribus will export only those pages of the document, which you have marked as “Ground Truth”
- For the export the program consults previous versions of your document. This means that if you choose to export all “In Progress” pages, the program will export all pages which have been marked as “In Progress”, even if their status is now updated.
- The program will export the latest “In Progress” version of your document.
- If you would like to export a specific former “In Progress” version of your page, open this version of the page in Transkribus. Open the Export window and select “Loaded version for current page”. In the “Pages” option, select “Current” before confirming.

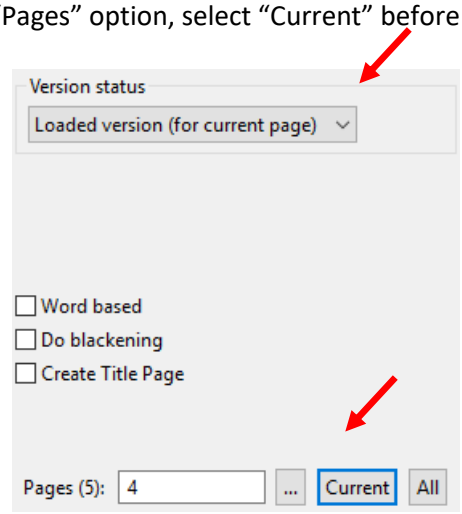

Figure 26 Export Loaded version (for current page)

Do blackening

- If you have blacked out sensitive sections of your transcription these words or phrases can also be hidden in the exported files.
- To do this, select “Do blackening” in the export options.
- **Note:** this option only works for Word, PDF and METS files.

Create title page

- Transkribus provides the option to create a title page. The page will be based on the information added in the “Document” tab within the “Metadata” tab.
- In “Document” tab you can add information about the title, author, language and date of your document. You can also create an Editorial Declaration to explain how exactly your document has been transcribed. For more on the Editorial Declaration, see How To Enrich Transcribed Documents with Mark-up.

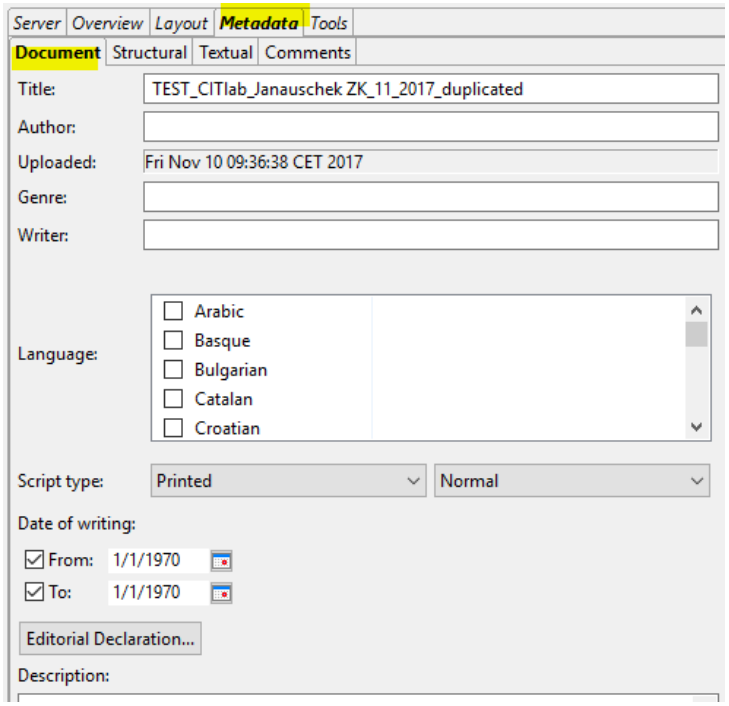

Figure 27 Information for title page

- This information will be included in your Title Page, as in the following figure.

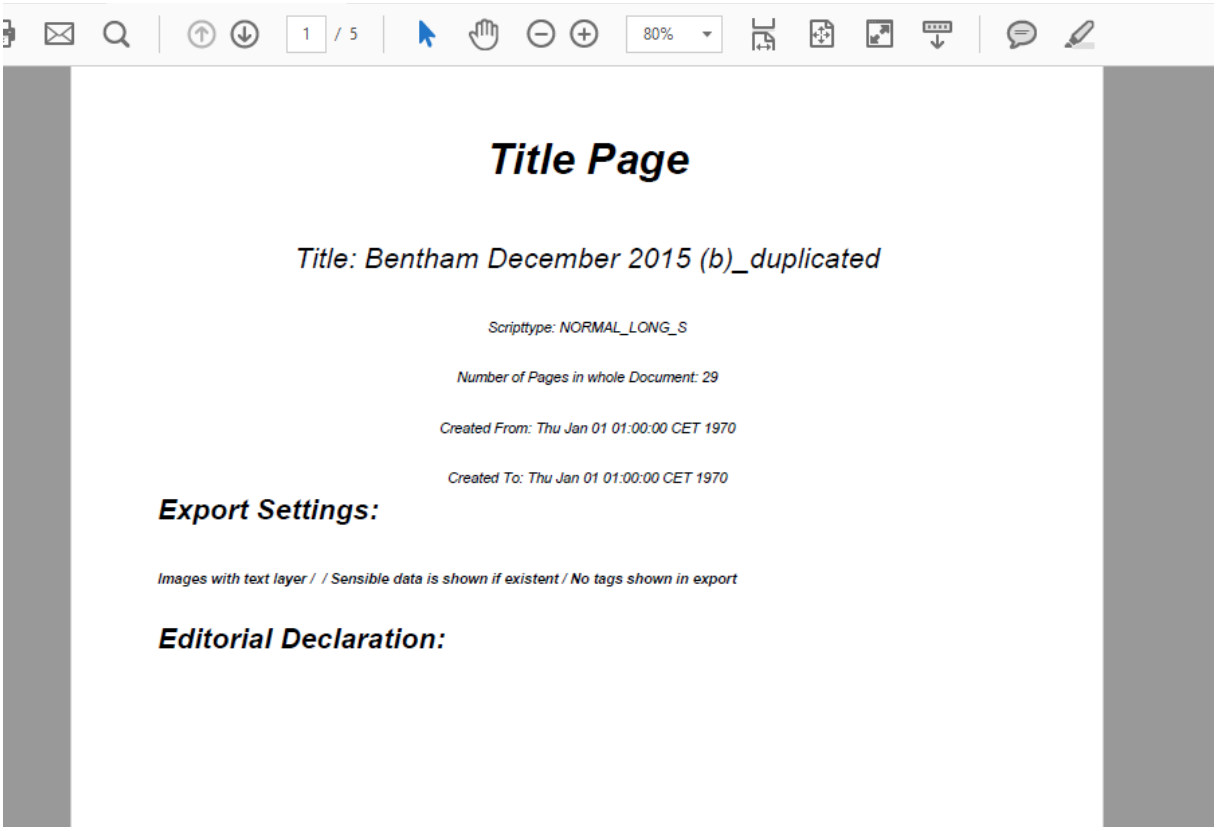

Figure 28 Title Page

Credits

We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission.