How To Transcribe Documents with Transkribus - Advanced Mode

This is a short introduction to the advanced steps for transcribing documents with Transkribus, software designed specifically to enable users to generate highly standardized transcripts. These transcripts can be used to train Handwritten Text Recognition technology and can also serve as a basis for scholarly editions.

For further information see the following papers and websites:

- How To Transcribe Documents with Transkribus – Simple Mode
- How To Prepare Test Projects with Transkribus – Especially for Archives and Libraries

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

- https://transkribus.eu/

Consult the Transkribus Wiki for further information:

- 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

Introduction ........................................................................................................................................... 3  
Transcription of documents – Simple Mode ......................................................................................... 3  
Segmentation ........................................................................................................................................ 3  
  Reading order ................................................................................................................................... 3  
  Reading order: Interline additions ...................................................................................................... 5  
  Reading order: Additions as extra notes ............................................................................................ 7  
Transcription ......................................................................................................................................... 7  
  Introduction ...................................................................................................................................... 7  
Diacritics and ligatures .......................................................................................................................... 7  
Punctuation marks .................................................................................................................................. 8  
Advanced mark-up ............................................................................................................................... 9  
  Tagging ........................................................................................................................................... 9  
  Predefined tags ............................................................................................................................... 9  
  Resolution of predefined tags ........................................................................................................ 10  
  Abbreviations and normalizations .................................................................................................. 10  
  Historical letters and abbreviation signs ......................................................................................... 10  
  Illegible text .................................................................................................................................... 11  
Deletions................................................................................................................................................ 12  
  Tagging entities: person names, geographical names, dates and similar entities............................. 13  
  Entities ........................................................................................................................................... 13  
  Predefined entities ........................................................................................................................ 13  
  Example .......................................................................................................................................... 14  
  Black out text ................................................................................................................................. 14  
Metadata .............................................................................................................................................. 15  
Editorial Declaration .............................................................................................................................. 16  
References ............................................................................................................................................ 16  
Credits .................................................................................................................................................. 17  

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

- In contrast to a “simple” transcription which mainly provides training data for the Handwritten Text Recognition (HTR) engines, an “advanced” transcription can be used for several other purposes. Namely it is suited for scholarly editions and also for e-learning applications targeted to help students and volunteers to “decipher” historical handwriting.
- For the advanced mode the following rules need to be taken into account:
  - The reading order of the text within the document may not always be linear. In the end you want to generate a correct “human readable” text from your transcription, so additions or marginalia will need to be handled carefully.
  - The transcription of historical texts may require the use of historical letters (e.g. to be found in Unicode Latin Extended character sets).
  - Abbreviations may appear frequently in the document and need to be expanded – again if human readable (and searchable) text will be available at the end.
  - Tagging of entities, such as persons, locations, dates, or works should be carried out in order to support easier access to the content of the document
  - Descriptive metadata and an Editorial Declaration need to be added to put the document into a wider context and to provide useful meta-information.

Transcription of documents – Simple Mode

- Download Transkribus and register for an account at the following website:
  - https://transkribus.eu/
- Once Transkribus has been installed, open the platform and click the ‘Login’ button in the Server tab.
- Log in using the email address and password you used when registering your account
- Before carrying on with this guide to advanced transcription, please read the following document:
  - How To Transcribe Documents with Transkribus – Simple Mode
- More information can be found on the Transkribus wiki

Segmentation

Reading order

- Once a document has been segmented into text regions, lines and baselines, you need to think about the reading order of the text.
- Many handwritten documents include corrections and additions added by the author, or someone else.
- In a scholarly edition you want to keep the reading order and maybe also express the fact that this text was an addition.
- For this purpose all segmentation elements can be ordered according to a user-defined order. This order can be displayed when selecting the “Item visibility” button in the Main Menu.
- The images below are taken from the Example Package: https://transkribus.eu/wiki/images/d/d6/Example_Package.zip
- For instructions on uploading the Example Package to Transkribus, see How to Transcribe Documents with Transkribus – Simple Mode

- The default reading order follows the topology of the text or line regions. All shapes are ordered according to the coordinates of the left corner of a text or line region.

- This mechanical reading order can be changed:
  - Click on the “Item visibility” button in the Main Menu, and you can then choose to show the reading order of text regions, lines, baselines or even words.
  - By clicking on one of the numbers marking the reading order, it is possible to type in a new number and change the reading order accordingly. The same can be done by moving the segmentation elements in the Layout tab.
Reading order: Interline additions

- Interline additions are a frequent way in which text is added to a document. For human beings it is simple to detect interline additions, but for machines this is a hard task.
- In order to generate the correct reading order, the following steps need to be performed manually:
  o Click the “Item visibility” button in the Main Menu
  o Select “Show lines reading order”
  o **Select the baseline below the addition** (if the addition is above the line).
  
  o Split the line region with the “Splits a shape with a vertical line” button in the Canvas Menu exactly where the addition should be logically placed.
- Edit the reading order so that it is correct: click on the number associated with each line region and then type the correct one.

Figure 5 Apply scissor V button to split the line region

Figure 6 Add correct reading order: 2 (first part of the line) becomes 1, 1 (=interline addition) becomes 2 and 3 (second part of the line) stays as 3.

Figure 7 Correct reading order after manual editing
Reading order: Additions as extra notes

- Additions which appear as extra notes (e.g. at the margins of a page) should be handled in a similar way to interline additions.
  - Note: Often such extra notes (or marginalia) are not part of the reading order but are “comments” and as such are on a different level to the primary reading order.
  - It will therefore be sufficient to mark them as “marginalia” in the Metadata tab.
- But if the extra note is really an addition to the running text and needs to be added in the reading order then it can be done in the following ways:
  - Option 1: The text region can be expanded so that all baselines of the addition are also part of the respective text region.
    - Note: You may use either rather large text regions, or you may use polygonal text regions. For this purpose select the “Add point to selected shape” button from the Canvas Menu.
    - Following the movement of your mouse pointer you can add points to the original text region and expand the shape in a way that it also includes the addition.
    - Afterwards the additional lines/baselines can be renumbered according to their correct reading order.
  - Option 2: You may generate just one large text region for the whole page and do the line/baseline segmentation manually in the correct order. In this way you will get the correct reading order right from the beginning.
    - Note: this may be the best option if you are dealing with a document which has a sophisticated layout with many additions, notes and deletions.
  - Option 3: You may connect the extra text region of the addition to the line where the text needs to be inserted. To do this, select both regions and then click the “Links two Shapes” button in the Structural tab, within the Metadata tab.
    - Note: The linking will be part of the XML file but is currently not supported in the export formats.

Transcription

Introduction

- A transcription which will serve as a basis for a scholarly edition should make more data explicit to the user and offer more contextual data than a simple transcription. In this case not only machine readability (i.e. training data for the HTR engine) but also human readability of the text will play an important role.

Diacritics and ligatures

- The correct transcription of diacritics and ligatures requires some expert knowledge. There are two main options for handling the correct transcription of these characters:
  - Option 1: Slight normalisation according to dictionary
    - The main rule to be applied here is the following: As long as you can clearly see the base character of a glyph and as long as the base character is also the one which is used in the dictionary to express this glyph, keep to the base character.
Example 1: LATIN SMALL LETTER Y will appear in many documents with an extra diacritical sign, indicating the history of this character coming from ii or ij. Therefore you find two dots or a something similar looking above the “y”.

Figure 8 German Kurrent Script: “bey”. Note: y is written as LATIN SMALL LETTER Y since the base character is still clearly visible.

In simple transcripts you will transcribe this as LATIN SMALL LETTER Y since the base character is clearly visible.

Example 2: LATIN SMALL LETTER S is expressed with two graphemes in most European historical scripts. We find therefore a clear distinction between LATIN SMALL LETTER S and LATIN SMALL LETTER LONG S.

Figure 9 “Thatbeſtand.” vs. “Revisionsgerichts”: LATIN SMALL LETTER LONG S vs. LATIN SMALL LETTER S

But although there is a clear distinction, a simple transcription would use LATIN SMALL LETTER S in both cases.

- Option 2: Palaeographic Transcription

Philologists or palaeographers are not only interested in the correct transcription, but also in the historical appearance and development of graphemes. Therefore it might also be interesting to transcribe the above examples with full support of the Unicode character set or even by utilizing the private area of Unicode.

Figure 10 Palaeographic transcription: Thatbeſtand vs. Kammergerichts

Note: Please take into account that this is an important decision and will affect the usability of the text in many ways. If you decide to go for a palaeographic transcription it will cause a lot more work than with a slightly normalized transcription.

Note: In printed texts (for which you can use Transkribus as well) the transcription of ligatures may play a role. Again the same rule can be applied: Though specific combinations of letters, such as “ft” are expressed with a specific grapheme where two graphemes are matched together, and though such ligatures can also be expressed with specific Unicode letters, we recommend transcribing them according to the dictionary.

Punctuation marks

- Punctuation marks are transcribed in the same way as characters. Use the appropriate character on your keyboard and do not normalize or add punctuation marks. Typical punctuation marks are:
- modern characters such as dot, comma, semicolon, colon: ".", ",", ";":"
- historical characters such as virgule (slash), or line fillers, etc.
- Note: Colons in historical texts are often used to mark abbreviated words. These should be transcribed as a colon.

- In contrast to many transcription rules where punctuation marks are added and omitted according to a modern understanding we recommend to keep to the original punctuation marks.
- If you want to add punctuation marks which do not appear in the original document you may use the “supplied” tag to indicate that the punctuation mark was added by yourself (see below).

Advanced mark-up

Tagging

- For adding meta information to the text, we have developed an annotation or tagging feature in Transkribus.
- You can find tags by clicking on the Metadata tab, and then the Tagging tab
  - To apply a tag you have to select text in the Text Editor field and afterwards click on the green + button of the tag you want to apply.
  - Note: Tags can be applied to text on region, line, word, or even character level. Users can apply as many tags as necessary to the text.

Predefined tags

- In order to simplify and standardize the transcription process we have already defined some tags and are supporting them with specific attributes. Tags and attributes are then used for encoding the TEI (Text Encoding Initiative) output files.
  - Note: You can always define your own tags and attributes, so that any kind of annotation can be handled with the Tagging system.
- The pre-defined tags should enable users to cope with the main operations necessary for an advanced transcription. More details are given below.
  - The following table gives an overview of all tags which we have foreseen for advanced text mark-up:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Purpose</th>
<th>Attributes</th>
<th>TEI Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbrev</td>
<td>Abbreviated words</td>
<td>expansion: expanded word/expression</td>
<td>&lt;abbrev&gt; (TEI)</td>
</tr>
<tr>
<td>Add</td>
<td>Indicate that text was added by the author or someone else</td>
<td>place: where the addition can be found</td>
<td>&lt;add&gt; (TEI)</td>
</tr>
<tr>
<td>gap</td>
<td>Gap: Omitted or illegible text</td>
<td>supplied</td>
<td>&lt;gap&gt; (TEI)</td>
</tr>
<tr>
<td>sic</td>
<td>Sic: Obvious errors</td>
<td>correction: for the correct form of the text</td>
<td>&lt;sic&gt; (TEI)</td>
</tr>
<tr>
<td>speech</td>
<td>Speech: verbatim text</td>
<td>speaker</td>
<td>&lt;speaker&gt; (TEI)</td>
</tr>
<tr>
<td>supplied</td>
<td>Supplied text by the transcriber or editor</td>
<td>reason: e.g. illegible text or</td>
<td>&lt;supplied&gt;</td>
</tr>
</tbody>
</table>
Resolution of predefined tags

The resolution of tags and attributes is performed according to XML and TEI conventions.

- Example "abbrev": The TEI export will handle the extensions in the following way:
  o `<choice><expan>Abbreviation</expan><abbr>Abbr.</abbr></choice>
  o Note: It is expected that the handling of common abbreviations will be standardized in Transkribus once a larger number of examples are available.

Abbreviations and normalizations

- Even more important than the transcription is the human readability of the text. Therefore the abbreviations need to be expanded as well. The abbreviated word needs to be tagged as “abbreviation” and the correct expansion needs to be written in the attribute field:

Historical letters and abbreviation signs

- In modern documents the handling of abbreviations is less important, but in historical documents it is a complex and challenging task.
- In earlier time periods words were often heavily abbreviated, in the hope of writing faster or saving paper. In some documents more than 20 or 30% of all words are abbreviated as shown in the figure below:
Again there are two main options to transcribe abbreviated text:

- **Option 1**: Keep to the rule which we already discussed above— as long as you can recognize the base character – **transcribe the base character**. This rule is especially suited to historians and people interested in the “content” of a document and those who want to provide training data for the HTR engine.
  - Therefore the correct transcription for the examples above would be simple:
    - pdor – qq – cus – qr
  - Note: In the future HTR engines may also learn to automatically expand these abbreviations (or to supply the correct abbreviation for an expansion) so that computer assisted transcription may be supported.

- **Option 2**: If you are also interested in using Unicode characters which are near to the special graphemes of the original document, then you can transcribe the text by utilizing the full power of Unicode.
  - In this case the transcription of above could look like the following:
    - pˁ: LATIN SMALL LETTER P COMBINING OGONEK ABOVE
    - ڈ: LATIN SMALL LETTER D WITH MIDDLE TILDE
    - o: LATIN SMALL LETTER O
    - 🫡: LATIN SMALL LETTER RUM ROTUNDA. Also LATIN SMALL LETTER R ROTUNDA may be used to represent this letter.
  - **Note**: In real-world cases it is often hard to decide which diacritic, modifier letter or Unicode character may be the right one. You may consult the MUFI website to get more information on this issue (cf. section “References”):
    - [http://folk.uib.no/hnooh/mufi/](http://folk.uib.no/hnooh/mufi/)

- **Mixed models will often be useful**. E.g. frequently occurring historical characters may be transcribed with their correct Unicode letter, whereas characters which were used just by a specific writer may be transcribed with their base character. You should note such “editorial decisions” in the Editorial Declaration so that your transcription rules are transparent to other users.
  - **Example**: LATIN SMALL LETTER RUM ROTUNDA 🫡 is regularly used in medieval and early modern texts. Therefore it might be useful to introduce this letter to an HTR model which deals exclusively with medieval documents and is dedicated to processing large amounts of such documents.

**Illegible text**

- Text which cannot be transcribed since it is illegible can be marked with the tags “unclear” or “gap”.
- Open the Tagging tab.
- Mark the text which is unclear in the Text Editor and tag it as “unclear” if it is hard to read or as “gap” if it seems impossible to read.

- You may also add alternatives or suggestions for the illegible word in the “Attribute” section of the tag.

Deletions

- If you discover deleted text you have several options:
  
  - **Option 1**: The text which is deleted is still readable, or at least large portions are readable. In this case transcribe the text as well as possible and mark it as “Strike through” from the Structural tab, within the Metadata tab.  
    
    - Note: HTR engines are able to also decipher strike through text and the more examples they have, the better.
  
  - **Option 2**: The text which is deleted is illegible, or only small parts can be read. In this case use the “gap” tag to indicate that there is some text which is illegible.
Tagging entities: person names, geographical names, dates and similar entities

Entities

- In order to make a document really valuable to a reader it is a common practice to create a register or index of entities, such as person and organisation names, dates and similar data mentioned in the text.
- It is also expected that Named Entity Recognition (NER) will become important in the future. HTR engines with their probabilistic output will support this task in an efficient way. But currently NER for historical documents delivers poor results – again the lack of training data is one of the most important reasons for this.

Predefined entities

- As with the advanced mark-up tags we have already defined some entities which might be helpful for a scholarly edition.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Purpose</th>
<th>Attributes</th>
<th>Tag used in the XML output</th>
</tr>
</thead>
<tbody>
<tr>
<td>blackening</td>
<td>Sensitive data which shall not be shown to a public audience</td>
<td>comment: an optional note why this text is not displayed to the public</td>
<td>&lt;blackening&gt;</td>
</tr>
<tr>
<td>date</td>
<td>Date: Any data concerning dates</td>
<td>year, month, day</td>
<td>&lt;date&gt; (TEI)</td>
</tr>
<tr>
<td>organization</td>
<td>Name of an organization</td>
<td></td>
<td>&lt;orgName&gt; (TEI)</td>
</tr>
<tr>
<td>person</td>
<td>PersonName: Names of persons</td>
<td>dateofDeath, dateofBirth, lastname, occupation, firstname, notize</td>
<td>&lt;persName&gt; (TEI)</td>
</tr>
</tbody>
</table>
Example

- In the Example Package you find tagged entities including “person”, “organization” and “date”.

  “7th April ’16” is tagged in this example as “date” and the attributes are filled in according to ISO standard: YYYY-MM-DD.
  
  Note: Currently there is no automation behind the standardisation of the date information, but this might be implemented during the READ project.

Black out text

- A very special tag is the “blackening” tag. It can be used to redact sensitive information in the export formats. Typically this is used to hide personal data in a document which is made publicly available.

- The blackening tag is completed with the “blackening” region which must be added with the segmentation tools. When you apply this tag, the transcribed text and the corresponding text in image (e.g. a person’s name) will both be hidden.
  
  Use the drop down menu on the “+…” segmentation element button on the Canvas Menu and select “Blackening”. Use the “Blackening” region to mark the word or section that you want to hide.
  
  Note: Click the “Item visibility” button on the Main Menu and select “Render blackenings” to display the blackened sections on a page.
  
  Mark a word in the Text Editor and select the “Blackening” tag. In the export of the document the text will be replaced by: [...].
Note: Within the export formats the actual image of the blacked out text (i.e. text covered by a black rectangle) is rendered, rather than just a layer appearing over the original image. This means that there is absolutely no way to derive the information from the blacked out segment of text.

![Figure 18 Sensitive data can be blacked out (will not be included in export formats)](image18.png)

![Figure 19 “Do blackening” to hide image regions and text in exported files](image19.png)

**Metadata**

- We are currently supporting only a very simple description of documents since we assume that in a Digital Edition most of the metadata would reside on an external server and be
linked to the document. Every document has its unique ID and can be accessed also via the REST services provided by the Transkribus platform (cf. Transkribus Wiki).

- The following fields are currently available:
  - Title
  - Author
  - Genre
  - Writer
  - Language
  - Script type
  - Date of writing
  - Description

Editorial Declaration

- Since there are always several ways to produce a correct transcript of a text it is important to describe the way “how” the transcription was undertaken.
- For this purpose we have included a special feature in Transkribus, called Editorial Declaration. This is found in the Document tab, within the Metadata tab.
- As with the Tagging system, the Editorial Declaration offers you a set of predefined features and options. Moreover you are enabled to create your own descriptions and to store them together with your document.
- It is especially important to list special characters and their use in the Editorial Declaration using the following form:
  - Character Set Extension: LATIN SMALL LETTER LONG S (U+017F)

References

To get an overview on scripts from Unicode: [http://www.unicode.org/charts/](http://www.unicode.org/charts/)

For historical transcriptions the following extensions are of interest:


- Contains e.g.:
  - Non-European and historic Latin
  - Phonetic and historic letters
  - Additions for Slovenian and Croatian
  - etc.

Latin Extended-C: [http://www.unicode.org/charts/PDF/U2C60.pdf](http://www.unicode.org/charts/PDF/U2C60.pdf)

- Contains e.g.:
  - Orthographic Latin additions
  - etc.


- Contains e.g.:
  - Medievalist additions
  - Insular and Celtic letters
  - Ancient Roman epigraphic letters
MUFI (Medieval Unicode Font Initiative)

- This initiative has collected and systematized about 1512 characters which are especially recommended for the transcription of medieval documents. Note: Some of them are still in the “private” section of Unicode, therefore not officially available.
- http://folk.uib.no/hnooh/mufi/
- http://folk.uib.no/hnooh/mufi/specs/MUFI-Alphabetic-4-0.pdf

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.