



Account ▼

<https://readcoop.eu>[Transkribus \(/transkribus/\) >](#)[How-To Guides \(https://readcoop.eu/transkribus/resources/how-to-guides/\) >](https://readcoop.eu/transkribus/resources/how-to-guides/)

Transkribus Transcription Conventions

# Transkribus Transcription Conventions

[🔗 Basic overview](#) [📋 Transkribus Expert Client](#)*Last update 2 years ago*[« How-To's overview \(https://readcoop.eu/transkribus/resources/how-to-guides/\)](https://readcoop.eu/transkribus/resources/how-to-guides/)

## ► About Transkribus ()

### Table of Contents



#### Transkribus Transcription Conventions

##### [Introduction](#)

##### Segmentation

##### Transcription

###### Diplomatic transcription

###### Combining words

###### Upper and lower case

###### Hyphenated words

###### Strikethrough text passages

###### Superscript text passages

###### Punctuation

###### Underlined text passages

###### Fonts

###### Additions and Reading Order

##### Conventions for special characters


###### Abbreviations

###### Diacritical characters


###### Equivalent of i/j or I/J

###### Equivalent of u/v or U/V

###### Ligatures



- Special characters
- Structural tags
- Account ▼
- Tagging
  - (<https://readcoop.eu>)
  - A few principles
  - Personal names
  - Tagging of abbreviations
  - Organisations
  - Dates
  - Gaps
  - Uncertainties
- Status
- Credits



This guide provides detailed transcription instructions for transcribing in Transkribus, providing guidance on features like abbreviations, diacritics and tags.

Users should move on to this guide after first consulting our basic transcription guide: [How to \(<https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/>\) Transcribe Documents with Transkribus – Introduction.](https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/)

(<https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/>)

## Introduction

Over the past few years, our transcription guidelines have been revised and simplified in line with our findings about what Handwritten Text Recognition (HTR) technology can learn to process about what Handwritten Text Recognition (HTR) can do.

This guide provides standardised instructions for transcribing historical documents in Transkribus.

Our aim is to help users produce transcripts quickly and efficiently, as the basis for strong HTR models that can recognise text with a high level of accuracy.

Users may want to produce a simple transcription that will only be used to train HTR technology to recognise their documents.

In this case, the most important consideration is to produce a consistent transcript that accurately represents the words in your document.

Alternatively, users may want to produce a rich transcription with additional tags and metadata that would be suitable for a scholarly edition.



(<https://readcoop.eu>)

This guide sets out common historical conventions for scholarly transcription that users can adapt according to their needs and the specific features of their documents. You do not necessarily have to change existing transcriptions that you have already completed. You may also find other effective ways of dealing with transcription issues that are not covered in this guide.

Transkribus users will also soon be able to transcribe documents in the Transkribus web interface, where it will be easier to transcribe documents in teams.

Before you begin, please check that you are working with the latest version of the Transkribus expert client:

- Click on the “≡” button and click “Check for updates”
- You can also try out the “Snapshot” versions, which are previews of the official releases of the platform

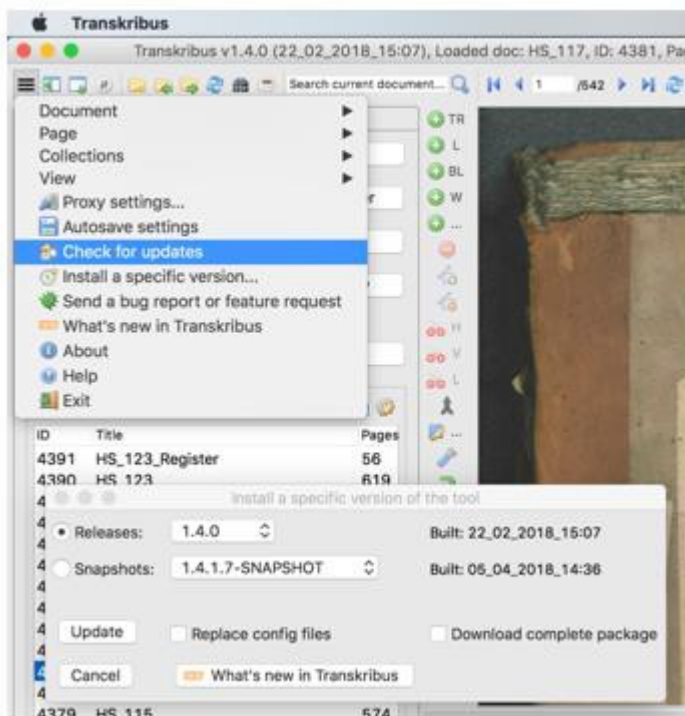


Figure 1 Checking for the latest version of Transkribus

## Segmentation

- In order to transcribe your documents in Transkribus, they must be segmented into text regions, lines and baselines.
- You can segment your documents automatically using the options in the “Tools” tab.



The baselines are the most important segmentation element

Baselines should always end flush with the line of text and they should not go outside the text region. (https://readcoop.eu/)

- The characters should “sit” on the baseline, any descenders should extend below it.
- If the layout of your documents is very challenging, you may need to perform some manual correction of the baselines.
- For more information on segmenting documents, consult: How to Transcribe Documents with Transkribus – Introduction.  
(https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/)

# Transcription

## Diplomatic transcription

The text needs to be transcribed character by character, according to what is shown in the image. Since uniform spelling rules did not exist in the past, orthographic and grammatical correctness is of secondary importance.

## Combining words

Words should be separated or combined according to the original text, even if it is not in accordance with the current practice.

## Upper and lower case

Again, the original text should be the basis for your decision. If an initial letter cannot be clearly identified as upper or lower case (majuscule or minuscule), the decision is up to you, but should be based on the current spelling rules.

## Hyphenated words

When hyphenated words appear at the end of the line, they should be transcribed and broken up

according to the original text. They no longer need to be marked with a “-” or a “tick”.

When hyphenated words appear in the middle of a line, they should be transcribed according to the original text.

(<https://readcoop.eu>)



## Strikethrough text passages

Passages of text that have been struck through should be marked up using the “Tag as strikethrough”

button in the Text Editor field.

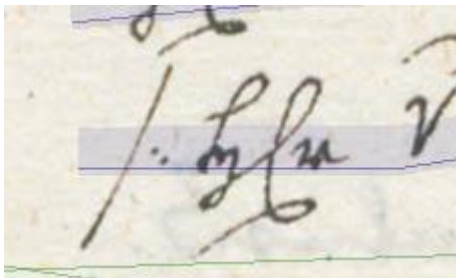
## Superscript text passages

Superscript text passages (including punctuation marks) should be marked up as superscript with the

“Tag as superscript” button in the Text Editor field.

## Punctuation

Punctuation should be transcribed using the keys on your keyboard, keeping as close as possible to the original.



**For documents of the 16<sup>th</sup> century and later:** Transcription should follow the original text, even if a punctuation mark has been used in a way that does not correspond to modern usage.

**For transcriptions of medieval texts:** Do not try to use modern punctuation. It would be better to either omit all punctuation or use specific symbols (i.e. Middle Dot, U+00B7).

Full stops often appear after numbers and headings – and they should be transcribed.

Sometimes historical documents use “/.” instead of brackets. In these cases, the “/.” should be

transcribed.

# Unread text passages



Account ▼



(<https://readcoop.eu>) Unread text passages are marked up with the “Tag as underlined” button in the Text Editor field.

## Fonts

Different fonts like Kurrent or Antiqua are not specially marked.

## Additions and Reading Order

Additions between the lines are segmented as separate lines and transcribed normally, but do not have to be specially marked. What is important is that the addition is placed in the correct place in the text according to the reading order. In such cases the reading order might need to be checked and revised.

The reading order should be as follows, according to the natural reading order of a human reader:

- Page number
- Header
- First section on top left
- First section on top right
- Etc.

To show the reading order options click on the “Shape visibility” button in the Main menu.



*Figure 2 Checking the reading order*

Select “Show baselines reading order” and numbers corresponding to each baseline will appear on your document image. Click on the number associated with a baseline to update its place in the reading order.

When interlineal additions appear, the correct reading order is: Text – Addition – Text.



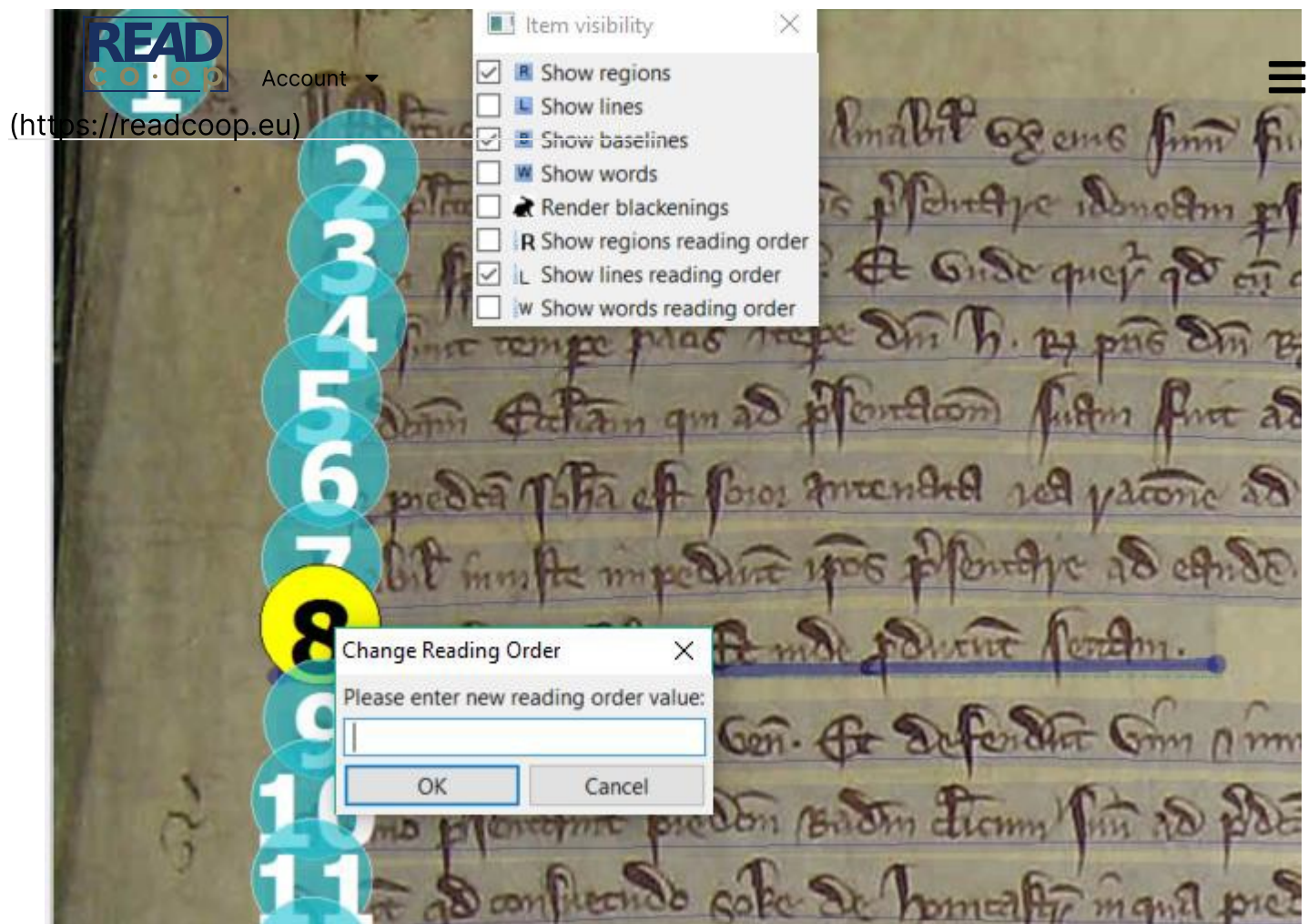


Figure 3 Changing the reading order

## Conventions for special characters

### Abbreviations

Normally abbreviations are transcribed according to the original text, i.e. not expanded. This includes historical abbreviations and abbreviations which are still used today (e.g. contemporary currency indications, titles and salutations). Only if the expansion involves only one or two additional letters it should be carried out.

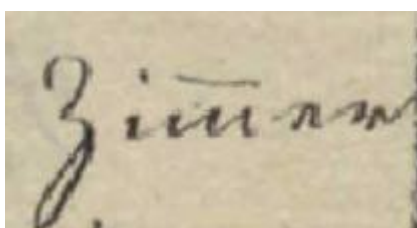


Figure 4. Example of an abbreviation: word with a nasal abbreviation on the m or n: Zi<sup>m̃</sup> er



Account ▼



(<https://readcoop.eu>)

## Diacritical characters

**Simple transcriptions:** diacritical characters (e.g. accents, circumflexes, cedillas, hyphens, tildes) can be ignored, except for the modern German umlauts.

**More elaborate transcriptions:** diacritical characters are transcribed according to the written characters on the page.

## Equivalent of i/j or I/J

The letters “i” and “j” can be used interchangeably. Again, the original text should help you make your decision. The two letters are transcribed as such, even if it does not correspond to the rules of modern orthography. Since they are often difficult to distinguish from each other (especially with capital letters), it is your own discretion or the spelling in use today that is decisive here.

**Alternative practice:** It is viable to only use “i” except for consonantal use of the letter.

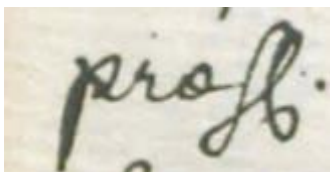
## Equivalent of u/v or U/V

The historical equivalents of “u” and “v” no longer exist because the letters are now used separately. Therefore, please adapt the transcription according to current usage.

Alternative practice: It is possible to use “u” and “v” as they would be read.

## Ligatures

Ligatures are common combinations of letters to form a new character.



“St” and “Sch” ligatures and the ligatures at the end of words or abbreviations should be transcribed in full. They do not need to be marked up as abbreviations.

For example, the ligature “præs” should be transcribed as “praes”.



The letter “s” can appear in different forms. Normal and long “s” (with descender) can both get transcribed as a normal “s” or according to their shape as “s” or “l” (U-017F). Double “s” or “ß” (sharp “s” or “Eszett”) are transcribed according to the original text.

## Structural tags

There are options for marking up the structure of your documents in the “Metadata”/“Structural” tab. The structure of the text is assigned at baseline level during the process of segmentation. You can mark up elements including page numbers, headings and marginalia and also train this layout with the P2PaLA-tool.

You can find more information about this in the How to Use the Structural Tagging Feature (<https://readcoop.eu/transkribus/howto/how-to-use-the-structural-tagging-feature-and-how-to-train-it/>) guide.

## Tagging

Personal names, places or locations, dates of various kinds as well as organizations, institutions or abstract identities can be marked with the corresponding tags.

You can find all of the tags in the “Metadata/Textual” tab.

**Note:** When it comes to HTR training, tags are not relevant yet. Developments in Named Entity Recognition technology should make the automated recognition of tags possible in the future.

For more information on enriching your documents with tags, see: How to Enrich Transcribed (<https://readcoop.eu/transkribus/howto/how-to-enrich-transcribed-documents-with-mark-up/>) Documents with Mark-up. (<https://readcoop.eu/transkribus/howto/how-to-enrich-transcribed-documents-with-mark-up/>)

## A few principles

- Please mark up only what is necessary; the characters and words that really belong to the appropriate tag.
- E.g. “d. d.” for “de dato” does not belong to the date itself and should therefore not be tagged as such.



Each tag should be applied separately to each word. If there are several (different) names or abbreviations next to each other, please tag them individually. Otherwise the search and normalization will not work.

(<https://readcoop.eu>)

- If required, several tags can be assigned to the same word, e.g. abbreviation, name, place, etc.

## Personal names

- When it comes to names, do not mark attributes (e.g. profession, origin, family, farm names, titles) before or after the name.
- EXCEPT: e.g. “Physikus Mr. XY”. In cases like this please mark everything as one name, because this is what is being referred to, even if the “Mr.” is in the middle.
- Also, words which refer to a unique single individual, but which do not include a name should be marked as persons (e.g. the mayor, the emperor, etc.).
- Indefinite terms like “the same” are ignored.
- If a name is mentioned at first and subsequently the person is referenced according to an impersonal label (e.g. “the baker”), then this indefinite name is not marked.

## Tagging of abbreviations

- Abbreviated words should be marked with the “abbreviation” tag.
- When two or more abbreviations appear consecutively, please mark each abbreviation with a separate tag.
  - E.g.: *Joh. Jak.*
- However, in the case of fixed phrases where two abbreviations appear consecutively, they can be marked up with a single abbreviation tag.
  - E.g. *d. d.* for de dato
  - *v. M.* (ultimo), *d. M.* (this month)
  - *I. J., k.k., p. C.* etc.
- If a word is abbreviated in any way you can either tag the whole word, even if only one letter is abbreviated at the beginning/end/in the middle, or you can tag the abbreviated part. Please be as consistent as possible.

## Organisations

- Everything is marked as organization or institution that is not an individual but nevertheless appears as a subject, agent or legal personality. Examples would be brotherhoods, offices or merchants.



Account ▼



- (<https://readcoop.eu>)
- Non-numerical dates that may not appear complete at first glance should be marked up, e.g. the Nativity of Mary, the month of September, the first quarter of 1792, etc.
  - However, please do not mark any periods as dates, e.g. three months.

## Gaps

- If the document is unreadable at any point due to difficult handwriting or strike through, the corresponding tag indicates this as a gap.
- Click your cursor where the illegible text appears and add the “gap” tag.
- If an illegible character or characters can be guessed, the relevant characters can simply be transcribed (without square brackets). Although it is common practice to add missing characters inside square brackets, this is unfortunately counterproductive when it comes to training the HTR engine.
- Supplied text, even single characters, should be marked up with the “supplied” tag.

## Uncertainties

- Any uncertainties can be marked with the “unclear” tag and, if possible, resolved later.

## Status

Please mark the edited or completed pages with the corresponding attribute in the status bar in the Main menu.

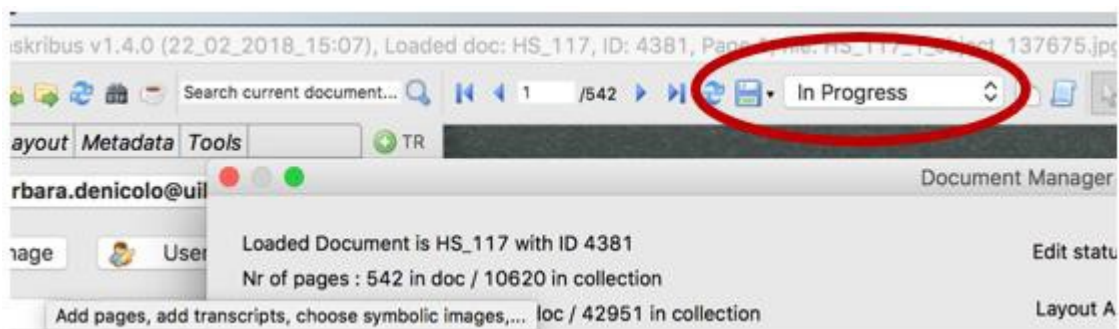


Figure 5 Defining the status of the document

The following statuses can be assigned:

**In Progress:** pages still to be transcribed

Done: pages that have been transcribed but which still need review.



Account ▼



Final: transcribed pages that have been reviewed as “Final”  
(<https://readcoop.eu>)

**Ground Truth:** transcribed pages that are completely finalised by the project administrator as

“Ground Truth” data suitable for HTR training. Once this status has been assigned to a page, it should

no longer be changed.

## 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 under grant agreement No 674943.

---

## The COOP

About us (<https://readcoop.eu/about/>)

Join us! (<https://readcoop.eu/join/>)

Our Members (<https://readcoop.eu/members/>)

Success Stories (<https://readcoop.eu/success-stories/>)

Work with us (<https://readcoop.eu/work-with-us/>)

## Products & Services

Transkribus (<https://readcoop.eu/transkribus/>)

Transkribus lite (<https://transkribus.eu/lite>)

Read&search (<https://readcoop.eu/readsearch/>)

ScanTent (<https://readcoop.eu/scantent/>)



Account ▼



## Useful information

[\(https://readcoop.eu/\)](https://readcoop.eu/)

News (<https://readcoop.eu/news/>)

Download Transkribus (<https://readcoop.eu/transkribus/download/>)

Public Models (<https://readcoop.eu/transkribus/public-models/>)

Payment and shipping (<https://readcoop.eu/payment-and-shipping/>)

## Helpful resources

Resource center (<https://readcoop.eu/transkribus/resources/>)

How-to Guides (<https://readcoop.eu/transkribus/resources/how-to-guides/>)

Getting started with Transkribus (<https://readcoop.eu/transkribus/start/>)

FAQs (<https://readcoop.eu/transkribus/questions/>)

Videos (<https://readcoop.eu/transkribus/resources/videos/>)



Copyright © 2023 READ-COOP SCE

[Terms & conditions \(https://readcoop.eu/terms-and-conditions/\)](https://readcoop.eu/terms-and-conditions/)

[Privacy Policy \(https://readcoop.eu/privacy-policy/\)](https://readcoop.eu/privacy-policy/)

[Contact \(https://readcoop.eu/contact/\)](https://readcoop.eu/contact/)

[Imprint \(https://readcoop.eu/imprint/\)](https://readcoop.eu/imprint/)

EN