Archiving Dossier Narrative
1st Medieval Latin Transcribathon from Polish sources
Launch Date: September 15, 2022
Archive Date: November 25, 2022

Current Project URL:
https://scriptores.pl/efontes/seminarium-efontes/transkribathon/

Launch Date: September 15, 2022
Archive Date: October 25, 2022

Abstract
The 1st Medieval Latin Transcribathon from Polish Sources crowdsourcing project was developed between September 15 and October 6, 2022. It consisted of the transcription of 11 medieval Latin documents from the Polish Crown Chancery Records in a short period of time by a group of volunteers with varying knowledge of Latin and experience in working on manuscripts. The transcription was carried out by the Transkribus platform, partly with the use of Unicode characters provided in the MUFI specification. The finished editions of the documents will be published on the *eFontes* project website, and will also become part of the *eFontes. Electronic Corpus of Polish Medieval Latin* while the transcriptions prepared during the project will serve as training material for the creation of the HTR models for medieval Latin documents.

Project Contributors

Primary organizers
Iwona Krawczyk
Jagoda Marszałek

Participants
1. Grzegorz Bartusik
2. Michał Kulisz
3. Mariusz Leńczuk
4. Agnieszka Maciąg – Fiedler
5. Bartosz Małecki
6. Ewa Orłowska
7. Kacper Pielat
8. Anna Poray-Wilczyńska
9. Fryderyk Rozen
10. Michał Słomski
11. Adam Talarowski
12. Orysia Vira
13. Tomasz Walczak
14. Jakub Zbądzki
15. Mateusz Zimny

Financial and Institutional Support
This project was supported in part by the *eFontes. Electronic Corpus of Polish Medieval Latin* project (Krzysztof Nowak, PhD) and, by extension, the Institute of the Polish Language at the Polish Academy of Sciences.

Narrative Section 1. Project Rationale and Scope
This project had three main objectives: the first was to obtain new texts of importance for Polish medieval researchers which would be fully searchable and in machine-readable format. The second was to prepare the

training material for the creation of the new HTR models for medieval Latin documents. The final goal was to create a group of old script enthusiasts in Poland in order to exchange knowledge about digital paleography and modern editing practices, as well as with a view to the possibility of establishing cooperation on future projects.
The main idea was to use a working methodology specific to citizen science in order to show how effective volunteer-based initiatives are, and what a help they can be when acquiring large amounts of data for various scientific projects. For this reason, all the tools used in the project had to be available in the public domain. The choice of source material was dictated by the relative scarcity of documents of this kind in the corpus of Polish medieval Latin. Finally, we selected 11 documents from the Polish Crown Chancery Public Register, from mid-15th century, which gave a total of almost 1,000 lines for transcription. The original manuscripts date from the reign of Casimir Jagiellon (1447-1454) and have been digitized by the Central Archives of Historical Records in Warsaw (AGAD) where they are also currently stored.

Communication
Communication was mainly via Slack messenger, in which participants exchanged questions and comments about the transcription on an ongoing basis. Information on the various stages of the work, meetings, etc., was also communicated via email. The Zoom platform was used for meetings and training. All instructions training materials were made available to project participants in the cloud on Google Drive.

Materials
The materials used during the training meetings were placed on a Google drive shared with the project participants. Among them was a comprehensive presentation on Latin paleography including a list of tools used in transcription work. An instructional video recorded for the project on how to use the Transkribus platform and a detailed discussion of the program's menu icons were also added here. Also included on the drive were Instructions on the characters used in transcription (prepared for the eFontes) and a detailed description of the rules for transcribing documents at each stage of the project.

Transcription tools and rules
We decided to carry out the project using the Transkribus platform. In order to do so, we created collections called Transkribathon, to which files with documents to be transcribed were uploaded. All participants were given permissions to edit each of the uploaded files. The transcription process was divided into two stages. In the first, i.e. transliteration, Unicode characters provided in the MUFI specification were used to represent abbreviations and selected characteristic letters of medieval manuscripts. In the second stage, the participants had to revise the transliteration made by their co-participants as well as transcribe the document in question with the development of abbreviations and partial normalization of spelling. An important part of the transcription process was also to improve the structure of the document, i.e. the lines of writing, as this is an essential factor for the subsequent building of the HTR model.

General transcription rules
Throughout the project, an effort was made to reproduce the document's graphics as similar to the original as possible. Accordingly, a number of principles were adopted:

In the first phase of the project (transliteration of a text):
- lower and upper case letters were not modernised
Note: In the case of the letter 's' at the beginning of words representing proper names (i.e. where capital letters can be expected, according to the rules of modern orthography), an upper-case S was written, in other cases a lower-case s.
- semivowels (v/u; i/j) were not standardised:
- abbreviations and selected letter variants were retained (according to the accepted character list)
- the notation of word spacing was not standardised
- no hyphens were inserted where words were moved to the second line
- punctuation was omitted.

In the second phase of the project (proofreading and transcription of the text):
- the spelling of lower and upper case letters was modernised (proper names, holiday names, names relating to divine persons - Deus, Dominus, beginning of a sentence, etc.)

- the notation of semivowels v/u, i/j was normalised: "u" and "i" for vowel position, "v" and "j" for consonant position: vni → uni, Iohannes → Johannes, iudex → judex
- abbreviations were expanded, no letter variant notation was used
- the word spacing was standardised
- no hyphens where words were moved to the second line
- punctuation omitted
- Roman numerals in dates were used in words and numbers
- mediaeval spelling was respected:
    - no diphthong notation used
    - "ci" before vowels was retained instead of the classic "ti": presencium
    - the notation of the letters "h", "w", etc. was not standardised
    - the letter 'w' was retained, also when replacing 'vu'→ nowm, ewm, inconwlsa
    - the letter "ÿ" was written in two ways:
    - "y" in place of the expected "y" or "i": dÿademate → dyademate
    - "ii" in the place of the confluence of two vowels "ii": attinencÿs → attinenciis
    - Polish words were written as close to the original as possible: "Wyelka Damba", "Malądamka"

Detailed transliteration instructions

The goal of the project was to prepare a transcription that would provide material for training an HTR model that would recognize types of abbreviations and preserve them in the transcript. It was very important that the transcription be consistent and coherent, so the set of Unicode abbreviations cited in the MUFI specification https://mufi.info/m.php?p=mufi was chosen. However, the selection was limited to the characters available to users without the need to install additional fonts, dispensing with private area characters (PUA), even if the shape or characteristics of the character were most similar to the symbol used in the transcribed documents (see the *Special characters* section) - in such cases, substitute characters were introduced.

I. Special characters

The table summarizes the Unicode characters used in the transcription, and gives a description of the character and examples from the documents under development.

A. single characters

ꝯ

To render the abbreviated prefix con-, com- at the beginning of a word (e.g. consensus), the Unicode character U+A76F (LATIN SAMLL LETTER CON) was used.

ɖ

For notation of the abbreviated endings of some pronouns and forms of gerunds and gerundives. i.e. -dem (e.g. quidem), -dam, -dum (e.g. legendum) written in the texts as 'd' with a dragged downward lasso, a similar Unicode character U+0256 (LATIN SMALL LETTER D WITH TAIL) was chosen.

ꝫ

To render the abbreviated ending -et in words such as qualibet or videlicet, the Unicode character U+A76B (LATIN SMALL LETTER ET) was used.

ſ

To render the ending -is, often after the vowels "c", "t", "d", "r", the Unicode character U+A76D (LATIN SMALL LETTER IS) was used. However, the meaning of this abbreviation was broader than that originally assigned to it (cf. Instruction - exceptions, etc.).

ʒ

The consonant "m" in word endings was written using the character U+0292 (LATIN SMALL LETTER EZH). This mark was also used in compounds (cf. Below).

ꝑ

Shortened syllables "per", "par", "por" at the beginning (e.g. partem) or in the middle of a word (e.g. corpore) were written using the Unicode character U+A751 (LATIN SMALL LETTER P WITH STROKE THROUGH DECENDER).

ꝓ

The Unicode character U+A751 (LATIN SMALL LETTER P WITH FLOURISH) was used to write the shortened syllable pro-, occurring at the beginning of a word (e.g., proinde).

ꝝ

The shortened syllable "rum" (sometimes also "ro") was written using the Unicode character U+A75D (LETTER SMALL LETTER RUM ROTUNDA). This character also appeared in other, less common functions (cf. Instructions - exceptions, etc.).

6

Due to the lack of a generally available Unicode character rendering the characteristic "s" at the beginning and end of words in the transcription, the numeral 6 was used in these ppositions. At the beginning of words that are proper or local names, this character was dropped in favor of the "S" in order to dispel doubts about the spelling of a lowercase or uppercase letter (cf. Editing rules).

ſ

The long "s" was written using the Unicode character U+017F (LATIN SMALL LETTER LONG S).

ẜ

The shortened syllables "sis", "ser" (e.g. servicia) were written using the Unicode character U+1E9C (LATIN SMALL LETTER LONG S WITH DIAGONAL STROKE). This character also occurred in other meanings (cf. Instruction - exceptions etc.).

9

The ending -us taking a form similar to the number 9 was rendered with this character. The Unicode character denoting this syllable is available in the private area in MUFI reccomendation(PUA).

ꝰ

The ending -us taking a form siimilar to the digit 9 written in the superscript was rendered with the Unicode character U+A770 (MODIFIER LETTER US).

ÿ

The double "i" in the short and long version (-ij) was written using the Unicode character U+00FF (LATIN SMALL LETTER Y WITH DIAERESIS).

B. compound characters

ⁱ

The superscripted "i", or the apostrophe-like symbol denoting it written in the superscript was rendered using the Unicode character U+0365 (COMBINING LATIN SMALL LETTER I).

̄

A horizontal dash (or tilde) above a letter, indicating the consonant "n" or "m" rendered using the Unicode character U+0304 (COMBINING MACRON). The character is placed above the letter followed by a consonant. This character also occurred as a general character and appeared in various functions and positions in the word (cf. Instruction - exceptions etc.).

̓

Similar in its shape to an apostrophe, the character replacing the consonant "r" or a syllable with this

consonant (e.g., "ri" in the word detrimentum) was written using the Unicode character U+0315 (COMBINING COMMA ABOVE RIGHT).

A shortened syllable with a consonant "r" occurring mainly in passive endings was rendered with the Unicode character U+035B (COMBINING ZIGZAK ABOVE). The symbol corresponding to this ending is available in the PUA.

p̄
To represent the abbreviated prefix pra-, pre-, -prae, the combination of the letter 'p' and the Unicode character U +0304 (p + COMBINING MACRON) was used.

qȝ
The abbreviated encyclitic -que was written using a cobination of characters: the letter "q" and the Unicode character U+0292 (q + SMALL LETTER EZH).

C. other

℞
The initial fragment of the word Reverendissimus etc. was written using the Unicode character U+211E (LATIN CAPITAL LETTER R WITH TAIL STROKE).

ꝙ
The abbreviated word quod (written in lowercase) was written using the Unicode character U+A759 (LATTIN SMALL LETTER Q WITH DIAGONAL STROKE).

Ꝙ
The same word Quod in uppercase was written using the Unicode character U+A758 (LATTIN CAPITAL LETTER Q WITH DIAGONAL STROKE).

II. Instruction – exceptions, etc.

Due to inconsistencies and discrepancies in the notation and placement of abbreviations, an auxiliary table has been prepared with examples from the text and transcription templates.
The table provides examples of, among other things, the transcription of letters or syllables in a superscript, the use of the character ¯ (U+0304 COMBINING MACRON), the use of individual characters applied in a function other than that originally assigned or located in a place other than that primarly accepted as typical for the character.

The characters used in the first stage of document transcription and included in the instructions were also included in the prepared Virtual Keyboard, which participants could add to the file from the Transkribus installation panel.

Narrative Section 2. Project Trajectory

Recruitment for the project was announced in early August 2022 on several SM profiles. Information about the Transcribathon has also been published on the eFontes project website (https://scriptores.pl/efontes/en/seminarium-efontes/transkribathon/). The deadline for accepting applications was set for September 5. Twenty-two people from several countries, i.e. Poland, Ukraine, the UK and Moldova, signed up for the project, eventually 15 people actively participated.
The project began on September 15 with a virtual introductory meeting, during which the project itself, the schedule and work division were presented and the principles of transcription were explained. It also provided an opportunity for the organizers to present the main objectives of the project and for all participants to get to know each other.

Next, two training sessions were held. The first was an introduction to Latin palaeography and short presentation of supporting materials (dictionaries of medieval abbreviations and Latin language, etc.). The second was devoted entirely to the use of Trankribus, and was also recorded for the convenience of project

participants. After the conclusion of the Trankribus session, the main phase of the project began: 2-stage transcription of documents.

In the first stage, which ran from September 16th to 23rd, participants were tasked with transliterating the documents (fig. 1, 2, 3) following the rules outlined previously.

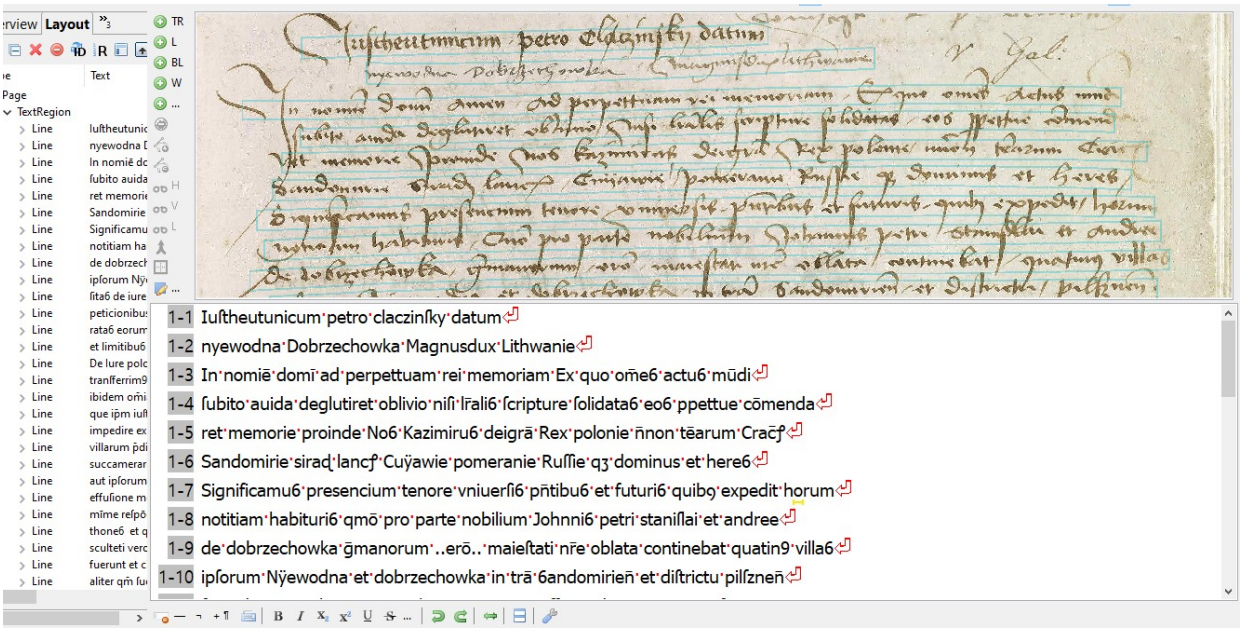

Fig. 1. Doc. 198 (page 1), 1st stage

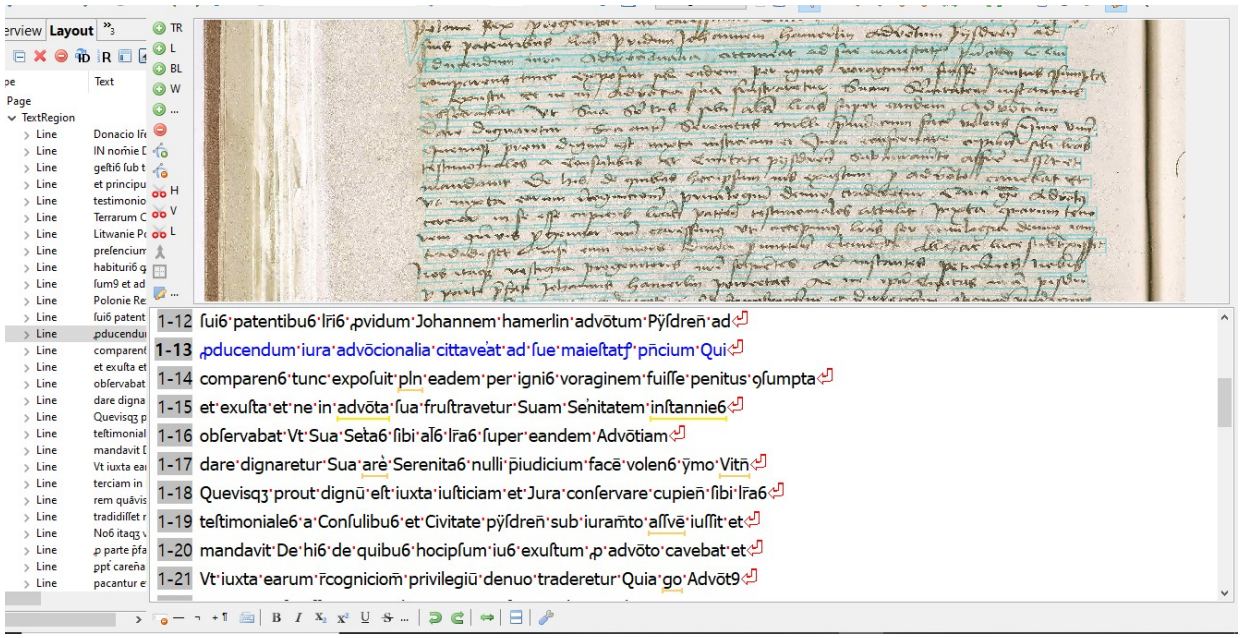

Fig. 2. Doc. 174 (page 1), 1st stage

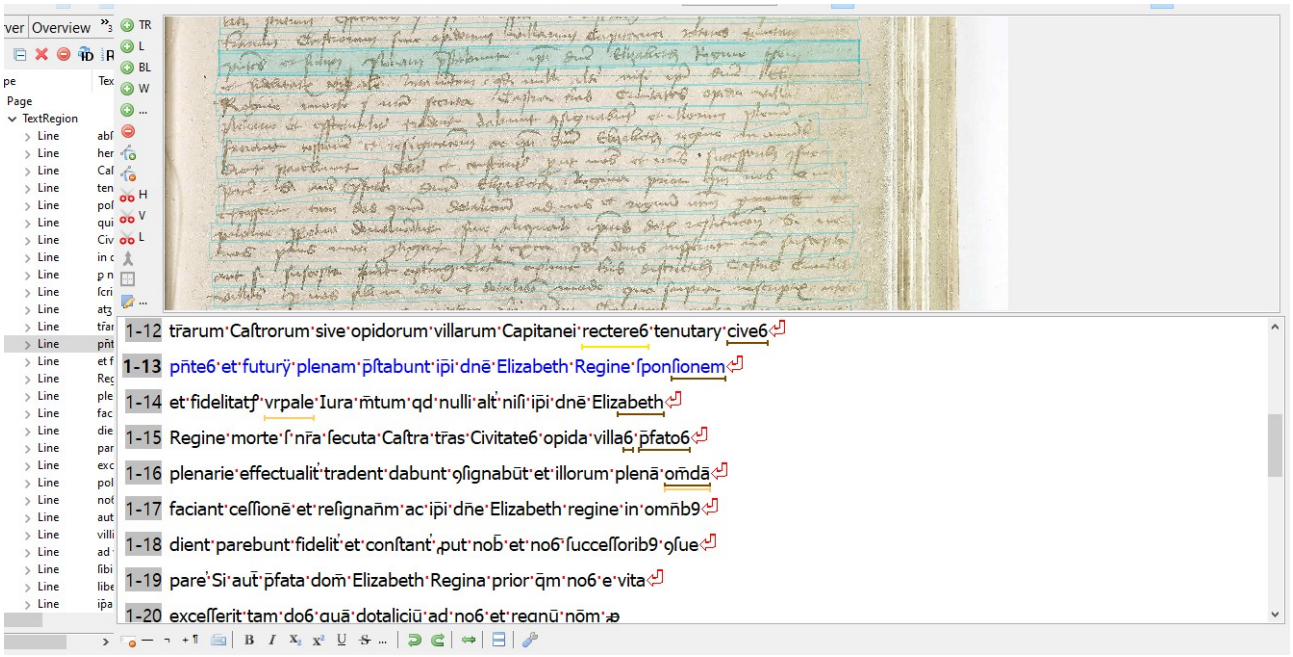

Fig. 3. Doc. 181 (page 2), 1st stage

In the second stage, which lasted from September 24th to 30th, participants had to revise the above transliteration (fig. 4) and provide a more traditional transcription, with expanded abbreviations and partial normalization of ortography (fig. 5).

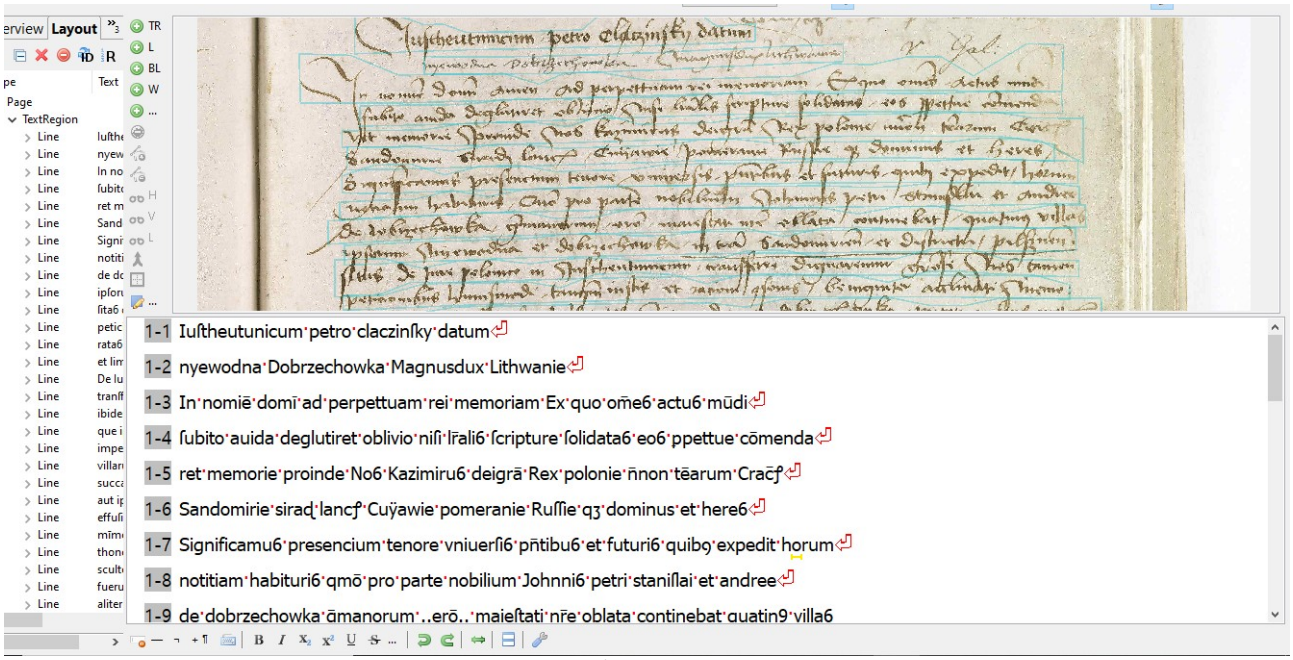

Fig. 4. Doc. 198 (page 1), 2nd stage – transliteration proofreading

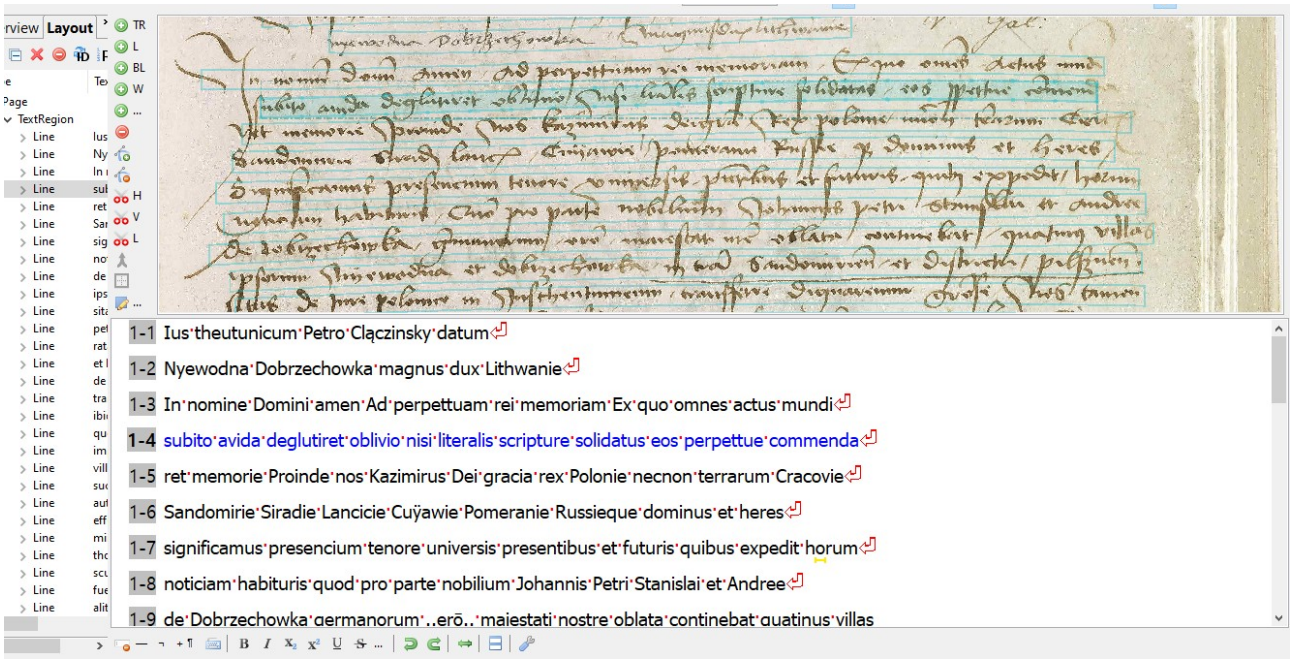

Fig. 5. Doc. 198 (page 1), 2nd stage – transcription

From October 1st to 6th, all developed materials (in transliterated and transcribed versions) were checked for compliance with editing instructions by the project organizers (fig. 6, 7).

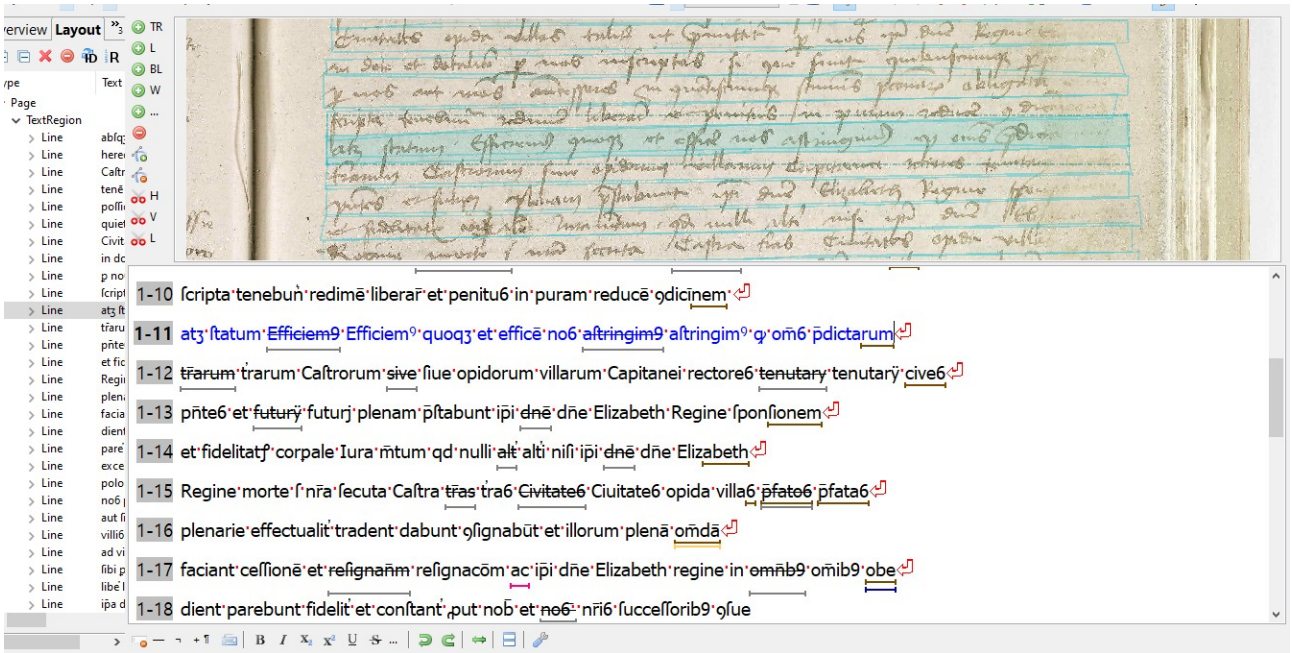

Fig. 6. Doc. 181 (page 2), 3rd stage – transliteration proofreading

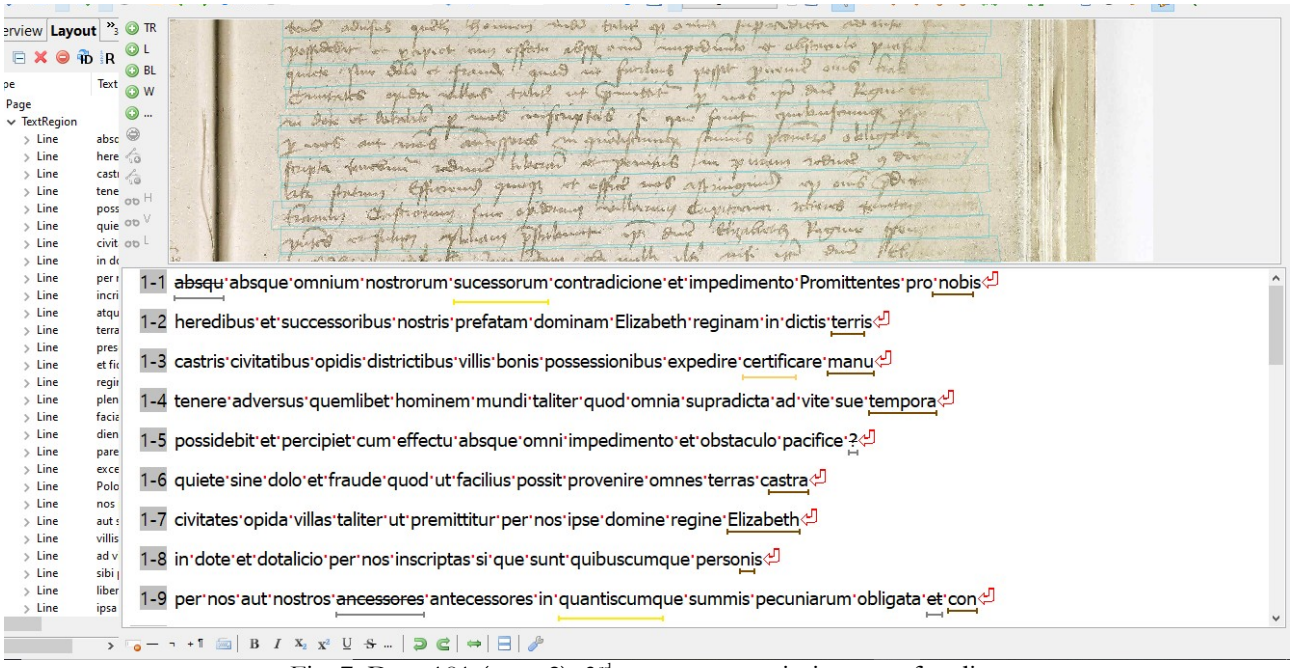

Fig. 7. Doc. 181 (page 2), 3rd stage – transcription proofreading

The results of the work after each stage were exported. Subsequent versions were uploaded to the server for the Transkribathon collection; at the final stage of the project, participants retained reader authorizations.

On the last day of the project, a closing meeting was also held, where the organizers presented the results of the Transcribathon and the participants were able to share their impressions of taking part in the project.

Narrative Section 3. Project-Specific Digital Objects and Platforms

The website for the project was created in Wordpress.com, as a subpage of the main eFontes project website (https://scriptores.pl/efontes/seminarium-efontes/transkribathon/).

Transcriptions of the manuscripts were created in the Transkribus platform, using 11 digitized manuscripts from the AGAD repository (http://agad.gov.pl/szukaj/?zbior=44&zestaw=0&data_od=&data_do=&q=&s=Poka%C5%BC+wszystko).

All documents for the team's use were created and made available in Google Docs or stored in a shared folder on Google Drive.

Digital objects collected for the project include the project website (converted to .pdf format), all project documents and all transcriptions provided by project participants, as well as a video - a presentation of the project made at the TUC 2022 conference.

Narrative Section 4. Project Outcomes (including analytics)

As a result of the project, it was possible to transcribe 10 of the 11 planned medieval manuscripts in only three weeks, the transcription of which will be made freely available for the public both as part of *eFontes. Electronic Corpus of Polish Medieval Latin*, and as separate editions published on the *Editiones Latinitatis Polonorum* project website. The material produced during the project, i.e. all transcriptions and transliterations will serve as a model for training the HTR models for Polish Medieval Latin.

The fruitful cooperation with the project participants inspires promise of future collaboration with the comminity of Latin palaeography enthusiasts going forward.

In addition, while the project was still in progress, it was presented at the Transkribus User Conference 2023 (September 29 - 30, Innsbruck, Austria, https://readcoop.eu/tuc22/), where it received considerable interest and resulted in proposals to establish collaborations with other researchers interested in techniques for automatic recognition of old handwritten texts.

Narrative Section 5. Documentation Statement

1st Medieval Latin Transcribathon from Polish Sources project has chosen the Attribution 4.0 International Creative Commons License.

Narrative Section 6: Project Bibliography
A full list of primary and secondary sources for this project, including references to projects that served as models for our Transcribathon, can be found in the project Supplementary Materials files.