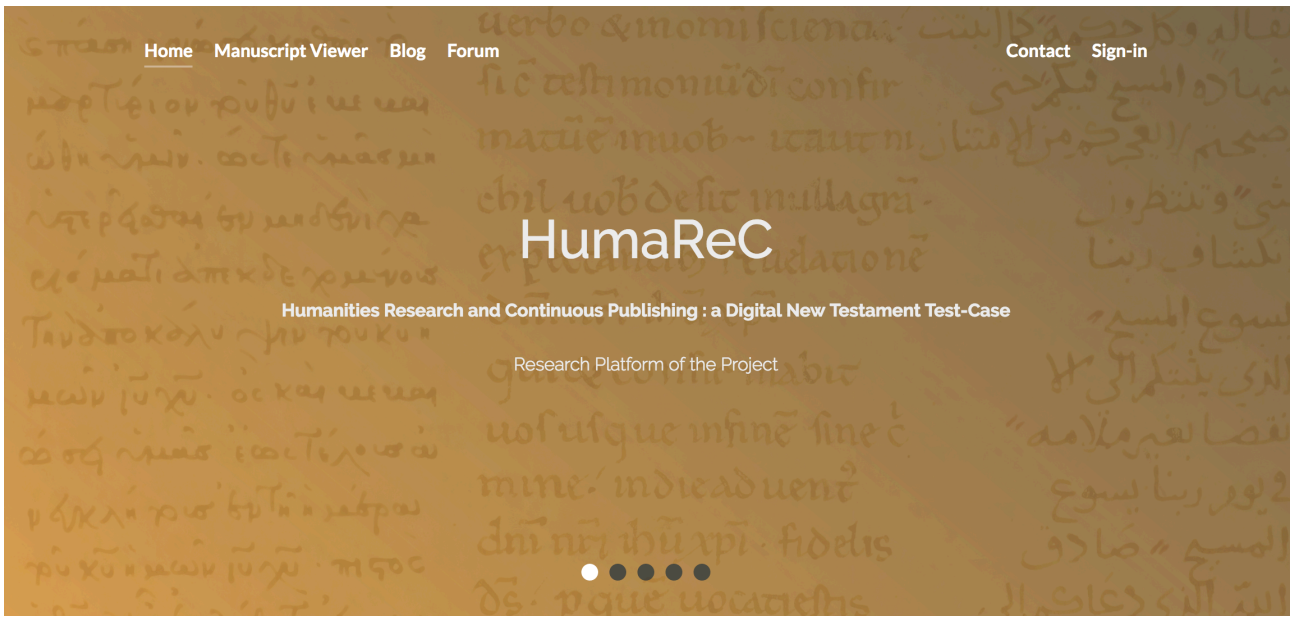
HumaReC – SNF project 169869, 2016-2018
Report #1 for the editorial and scientific board



Revision History

Revision	Date	Author	Organisation	Description
0.1	21.06.17	S. Schulthess	SIB	Outline
0.2	26.06.17	C. Clivaz	SIB	Revision
0.3	27.06.17	M. Sankar	SIB	Revision
0.4	29.06.17	A. Chasapi	SIB	Revision
0.5	until 20.07.17	Editorial board	International	Revision
1.0	21.07.17	S. Schulthess and team	SIB	Final version

AC: Anastasia Chasapi, CC: Claire Clivaz, MS: Martial Sankar, SS: Sara Schulthess

Table of Contents

I. Research Platform 2
I.1 Homepage 2
I.2 Manuscript Viewer 2
I.2.1 Technology and visualization 2
I.2.2 Transcription and encoding 3
I.3 Research blog 4
I.4 Forum 5
I.5 Annotation tool 5
I.6 Web book 5
II. Research on the manuscript 6
II.1 General 6
II.2 The text 6
II.2.1 Relationship between the three texts 6
II.2.2 Origin of each text 6
II.3 Experience with Transkribus 7
III. Development of a new DH research, editing and publishing model 7
IV. Communications 8
V. Perspectives 9

I. Research Platform

During the first months, AC, MS and SS worked mainly on the creation of the research platform. After consideration of a number of options and with particular attention to security and flexibility, the team decided to create the website using the Joomla content management system. The platform, which was set up by AC and MS with the input of SS and CC, provides a homepage, a manuscript viewer, a research blog and a forum. In the short term, it will also provide a redirection to the web book.

The research platform is hosted at the Vital-IT servers. AC and MS are in close collaboration with the Vital-IT system administrators to ensure the website's accessibility and security.

The platform was launched on March 7th 2017 and is accessible in full open access at https://humarec.org.

I.1 Homepage

The Homepage (https://humarec.org) provides information about the project and about the team. It also redirects the visitor to the other sections of the website: manuscript viewer (I.2), blog (I.3), forum (I.4), and web book (I.5).

I.2 Manuscript Viewer

I.2.1 Technology and visualization

HumaReC provides a manuscript viewer (http://humarec-viewer.vital-it.ch) in full open access with quality images and many features for the study of the manuscript.

The manuscript viewer is based on the EVT technology.[1] EVT is an open source software and the

[1] Edition Visualization Technology. Website of the tool: http://evt.labcd.unipi.it/; website of the Digital Vercelli Book project: http://vbd.humnet.unipi.it/beta2/; see also the publication: Roberto Rosselli Del Turco et al., 'Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions', *Journal of the Text Encoding Initiative*, no. 8 (2014), http://jtei.revues.org/1077. All links were last accessed 20/06.17.

experiences made with EVT for the *Tarsian* edition were very positive.[2] The tool EVT is designed to create a manuscript viewer with a user-friendly interface from XML TEI-encoded texts.

We considered several options for the display of the three languages and, using the advantages of the EVT tool, have for now chosen to display the several edition levels. The user can therefore select in the scroll menu between Greek, Latin or Arabic in different transcription options.

Figure 1: Display option

Another option (since the release of EVT 1.2) would have been to use column breaks:

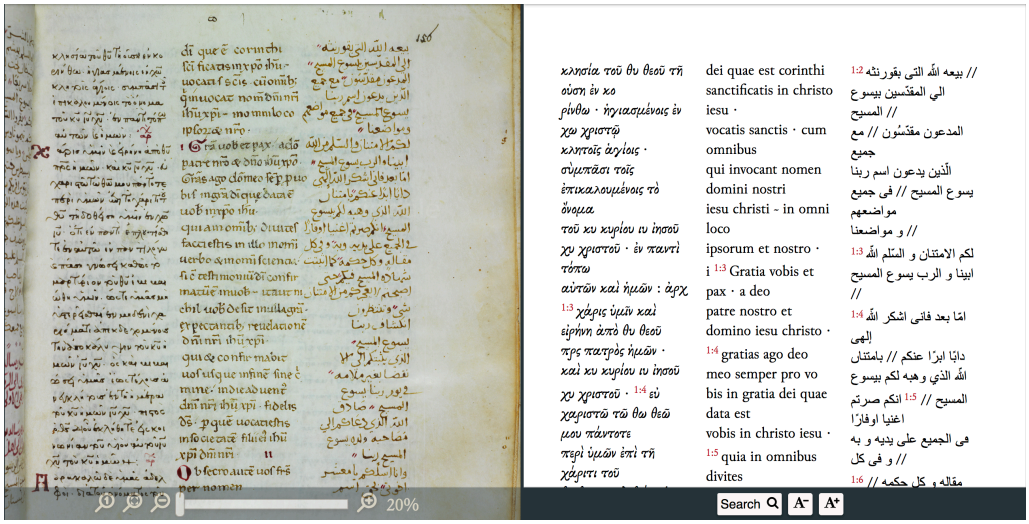


Figure 2: Layout for three columns

A large computer screen is needed for the optimal display of such a layout as in Figure 2, however, and in order to offer a good readability for every user we stayed with the first option (Figure 1).

I.2.2 Transcription and encoding

In continuous publishing, new folios are directly made available when they can be uploaded. Two stages, transcription and encoding, are needed for the integration of text to the manuscript player. The text is transcribed and afterwards encoded according to the XML TEI standards. This makes the text readable by the software EVT with the basic TEI heading and page break mark ups, and allows adding additional

[2] https://tarsian.vital-it.ch. See: Claire Clivaz, Sara Schulthess and Martial Sankar, 'Editing New Testament Arabic Manuscripts on a TEI-base: fostering close reading in Digital Humanities'. *Journal of Data Mining and Digital Humanities*, Episciences.org, 2017, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, https://jdmdh.episciences.org/paper/view?id=3700.

information such as the line breaks, the color of the text, and separation into verses.[3] The encoding also makes it possible to clarify the choices made during the transcription and to let the users choose which levels of transcription they are interested in:

- the 'diplomatic' version corresponds to a transcription of the text that follows the spelling and other characteristics of the manuscript as closely as possible.
- the 'normalized' version corresponds to a text that has been normalized and standardized for the reader. For the Greek, we have written out the *nomina sacra* and added some accents (mostly breathings); for Arabic, we have added diacritical points when they were missing or misplaced.[4]

Here is an example of how the two levels concretely work:

> We read in the Greek column the noun David that is abbreviated as δα͞δ; we expand this in δαυὶδ to help the reader with this abbreviation. This is encoded as such:
> `<choice><orig><am>δ&#x0360;αδ</am></orig><reg><ex>δαυὶδ</ex></reg></choice>`
> If readers choose to display the diplomatic version, δα͞δ will appear; if they choose the normalized version, δαυὶδ will appear. The alternative is in both cases displayed when the user run the mouse over the word:

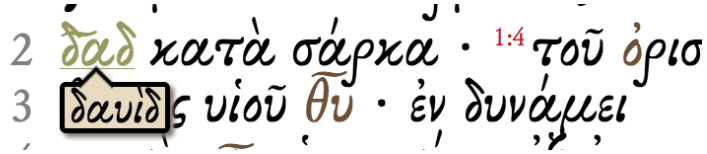


Figure 3: Example 'diplomatic' / 'normalized'

The addition of information for the text-image linking at the line level also takes place during encoding, a function included in EVT (Text-Image button). The description of the line surfaces with coordinates was made manually with the help of the Image Markup Tool (IMT), a task that is time consuming. This could be facilitated by the use of the Transkribus tool (see II.4).

SS has transcribed the three columns of 46 folios and encoded 20 of them that are now available online.

For now, SS has chosen not to invest time in the diplomatic Latin version. In fact, due to the diversity and complexity of the abbreviations and the use of special characters, the elaboration of a diplomatic text for Latin is more complicated and much more time-consuming than for the Greek and the Arabic. We hope to have time during the course of the project to work further on encoding the Latin text.

I.3 Research blog

The research blog (https://humarec.org/index.php/continuous-publications-blog) is part of the continuous publishing process. In the blog, we offer regular updates about the development of the project and the research results. Four announcements and ten articles are online for now.

The articles seem to reach people (according to Facebook likes, hit views, etc.); reactions in the form of comments, however, have been limited. Since we are interested in feedback, strategies regarding exchanges with readers will be further enhanced (see also I.4 et I.5).

---

[3] XML files are available on github: https://github.com/humarec/humarec-tei.

[4] This is explained in the blog article 'How to Use the Manuscript Viewer', https://humarec.org/index.php/continuous-publications-blog/11-articles/16-use-manuscript-viewer.

I.4 Forum

The forum (https://humarec.org/index.php/forum) was set up in order to facilitate communication with people interested in the project and was integrated in the website using Kunena, a Joomla forum plugin.
Even though we have received great input,[5] we are not yet satisfied with forum activity. We would like to receive more reactions from users about the manuscript and the transcription. This is why AC and MS are working on the implementation of an annotation tool (see I.5).

I.5 Annotation tool

The annotation tool aims at allowing the users to directly comment on the manuscript viewer. Prior to development, AC and SS investigated existing text annotation options. The tool should be compatible with the manuscript viewer. Several noteworthy solutions exist, varying from simple customizable Javascript libraries (for example Annotatorj - http://annotatorjs.org/) to complete out of the box solutions (for example Hypothes.is - https://web.hypothes.is/). In the case of HumaReC we wanted to be able to have a simple tool that would permit us to store the annotations in a local database hosted in the Vital-IT servers and therefore allow us to moderate the manuscript inputs. Based on this reasoning, AC is currently developing an annotation tool using Annotatorj at its core.

We hope that this tool, which is more intuitive for the users, will encourage them to comment the transcription more often.

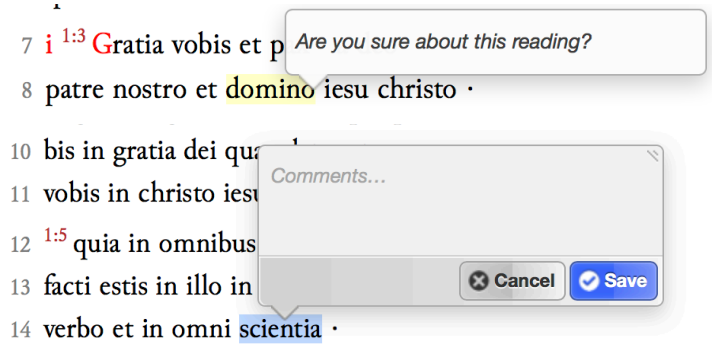


Figure 4: Annotation tool

I.6 Web book

The web book will summarize the research made during the project. Similarly to a conventional book, this text will be long, structured, and published by a publisher, but the digital medium will also offer additional, enhanced features. The web book will be written continuously, allowing the readers to follow the progression of the research and allowing their feedback to be included. Furthermore, the book will be related to other data on the platform, such as to the manuscript viewer, enabling users to move readily between the resources.[6]

The publisher Brill is very interested in this new format; a first announcement about the publication will be included in the next Brill catalogue. As we are planning a partnership with Brill, we chose to present the web book in a simple and flexible structure, which could at any time become independent from our research

[5] One user comments regularly the new folios transcribed, see https://humarec.org/index.php/forum/transcription.
[6] The principles of the web book are explained in a blog article: 'Web Book', https://humarec.org/index.php/continuous-publications-blog?start=10.

platform. Thus, we are developing the web book from an open source book made by Google.[7]

The web book is not yet available in open access but can be accessed here for the editorial board, with a password: https://humarec.org/webbook/book/index.html.

II. Research on the manuscript

II.1 General

While preparing the transcription, SS also engages in research on the text by comparing the columns with one another and also with other Greek, Latin and Arabic witnesses. In addition, she investigates the historical context of the manuscript.

In the blog article 'About Marciana Gr. Z. 11 (379) and its relatives'[8], SS published important general remarks about the similarities of our manuscript and two other manuscripts, the trilingual Psalter Mus. Brit. Harl. 5786 and the Gospels Greek-Arabic manuscript Marciana Gr. Z. 539 (303). This closeness has led us to suspect that the Marciana Gr. Z. 11 (379) should be considered as a 12th century manuscript as well. A 12th century date would place the manuscript's origin during the Norman domination of South Italy, including Sicily, which would fit with the trilingual feature of the manuscript.

As a next step, paleographical analysis of the scripts may strengthen the proposed 12th century dating.

II.2 The text

II.2.1 Relationship between the three texts

The texts of the three columns are independent: unsurprisingly, the Latin is not translated from the Greek but is a copy of a Vulgate text. Likewise, the Arabic text is not a translation of one of the columns. The translation is of Latin origin, from a text of the Vulgate but not of the same version as found in the Latin column. At several points, however, the three traditions have influenced each other – directly during the writing process or through the liturgical practices that were in place at the time. These observations are presented in the blog article 'Three (in)dependent versions of one text' on the basis of several examples of interaction between the three versions.[9] A second article, 'Where did Cilicia go', highlights this phenomenon with a particular reading in Gal 1:21.[10]

II.2.2 Origin of each text

The Greek text (manuscript 460 in Gregory-Aland list) is close to the Byzantine text. We support this statement with an analysis of 46 readings and will extend our study to the rest of the published folios.[11] Though the Greek text is near to the Byzantine text, we also observed several uncommon readings (see blog article); however, these readings are probably due to an influence of the Latin tradition (see point above).

The Latin text is a copy of the Vulgate. Our work with the Latin text seems to indicate a tradition close to the one transmitted in the Sixto-Clementine edition; however, a close analysis of this point is required.

As we said above (II.2.1), the Arabic text is of Latin origin, but several differences with the Latin column

[7] https://github.com/captn3m0/google-sre-ebook.
[8] https://humarec.org/index.php/continuous-publications-blog/11-articles/21-about-marciana-gr-z-11-379-and-its-relatives.
[9] https://humarec.org/index.php/continuous-publications-blog/11-articles/29-three-independent-versions-of-one-text.
[10] To be published soon.
[11] Blog article 'First look at the Greek text', https://humarec.org/index.php/continuous-publications-blog/11-articles/23-first-look-at-the-greek-text.

have led us to conclude that the *Vorlage* is a different text. The fact that the translation was made from Latin is very exciting because only two other New Testament translations from Latin into Arabic are known. One of them is Vat. Lat. 12900, a very old fragment of Galatians that goes back to the 9th century. The second is Madrid BN 4971. A comparison between the three manuscripts – Marc. Gr. Z. 11 (379), Vat. Lat. 12900 and Madrid BN 4971 – shows that they are part of one same family.[12] Marc. Gr. Z. 11, which is almost complete, is a very interesting new link in the study of New Testament Arabic translations of Latin origin. This find also sheds light on the influence of Mozarab culture in Sicily, as Vat. Lat. 12900 and Madrid BN 4971 are of Spanish rather than Sicilian origin. This Mozarab heritage is also an important element in the occasional use of unexpected vocabulary, an aspect of the project that invites further study.

II.3 Experience with Transkribus

Transkribus is a Handwritten Text recognition tool.[13] Unlike OCR (Optical Character recognition), the technology does not focus on individual letters but it processes the images of entire lines. This technology is much more adapted to handwriting. In order to train the machine for a specific handwriting, a certain amount of folia must be transcribed manually in a first phase.

We performed a first test for Greek and Arabic with 5,000 words for each language. As the dataset was small, the results were low, with more that 30% of Character Error Rate. As soon as 15,000 words are ready, we will perform a second test. Since we are working with a small amount of folia, it is uncertain if Transkribus will help us with the transcription work, but working with the tool will lead to interesting results in other projects.

We have also been employing the XML exports that are offered by Transkribus. In fact, it would be very useful if we could use directly the TEI XML files exported from Transkribus with the technology of EVT. Transkribus automatically recognizes text lines and produces coordinates for the line surfaces that are included in the TEI XML files. We would gain time if the files could be directly readable by EVT, as the lines have to be defined manually for now (see I.2). SS has collaborated with the Transkribus team on this point[14] and we are not far away from concrete results.

III. Development of a new DH research, editing and publishing model

CC has achieved several important steps in the building of the new research, editing, and publishing model.

Each important step is checked with the SNF division 1: it belongs to the research project to build this new model with diverse partnerships.

The collaboration with the Marciana Library was a first important step. After looking at the platform HumaReC, the Marciana library proposed a fair price for permitting the images to made available online in open access on the HumaReC platform. Moreover, the Marciana Library and Transkribus have now signed a Memorandum of Understanding that allows our team to fully use the images in the Transkribus tool. Our evaluation of the efficiency of Transkribus for transcribing a specific manuscript will add important information for future collaborations.

---

[12] See blog article 'Important comments on the Arabic translation', https://humarec.org/index.php/continuous-publications-blog/11-articles/25-important-comments-on-the-arabic-translation.

[13] Transkribus is a Research Infrastructure funded by the H2020 Project READ - Recognition and Enrichment of Archivial Documents, https://transkribus.eu/Transkribus/.

[14] See the assigned issue on the Transkribus Github: https://github.com/Transkribus/TranskribusSwtGui/issues/40.

Next, a publishing model was negotiated between Swiss actors (SNF and BNS) and the international publisher involved in the project, Brill. In May, the HumaReC platform received an ISSN from the Swiss National Library (ISSN 2504-5075): all the published material associated with the project can be referred to with this number. HumaReC can now be found in the Swiss library catalog Helveticat: http://www.helveticat.ch/lib/item?id=chamo:1893712.

Another ISSN will be given in due time by Brill to the Web Book, after approval following the final peer-review. The construction of the Web Book model is at present a more challenging point. We are convinced that the writing of a long text – our former monograph paper – belongs to the Humanities core skills. New research results are made clear only by the slow and attentive exercise of the writing, yet, through the digital media, this writing can of course be multimodal and welcome multimedia material.

Brill has confirmed its interest in this format and has offered to insert the project in the next Biblical studies catalogue in August. The Web Book will be progressively published and completed in a pre-peer-review phase on HumaReC and will stay related to this ISSN until its completion. In fact, remarks and discussion are already welcomed. When completed, the book will be peer-reviewed in a process led by Brill, as is the procedure for a usual monograph. After acceptance, the project will receive its own ISSN and become a full Brill product. At the computing level, the Web Book is an independent object and a 'dump' can be provided to Brill. Following this initial attempt the possibility to develop this model at a middle or larger scale could be a next step, with an easy editor for SHS researchers.

Last but not least, an international editorial board is already providing input back about HumaReC (via email exchanges). This report to the editorial board will be a formal step in the consultation. We hope to receive more regular feedback about the next developments in the coming weeks.

IV. Communications

In addition to the platform, the project is active online through a Facebook page (facebook/humarec) and a Twitter account (@project_humarec). The Facebook page has 142 followers, which is satisfying for a project of this kind. The Twitter account has "only" 24 followers but the tweets are often retweeted by popular accounts, which gives us a good visibility.

SS has presented our research at several conferences. She presented the HumaReC project at the 3rd EADH Day in Roma (January 25th 2017) in benefit of an EADH bursary for attending the meeting.[15] She was invited to speak at the *Translators, copyists and interpreters: Jews, Christians and Muslims and the transmission of the Bible in Arabic in the Middle Ages*, where she presented the discoveries about the Arabic text (II.2.2) (April 27th 2017).[16] She also presented the project and some aspects of the text in Zurich (May 13th 2017) at the *Neutestamentliches Kolloquium der schweizerischen theologischen Fakultäten*.

SS will attend as guest the 72th SNTS Meeting in South Africa and will deliver a paper in the New Testament Textual Criticism seminar: *Marciana Gr. Z. 11 (379) and the Project HumaReC: An Innovative Research Model for an Unique New Testament Manuscript* (August 9th 2017).

[15] http://aiucd2017.aiucd.it/?page_id=996.
[16] http://www.uco.es/servicios/publicaciones/ocs/index.php/translators_copyists/.

CC will present a poster at the DH 2017 in Montreal, prepared by all the team: Sara Schulthess, Anastasia Chasapi, Ioannis Xenarios, Martial Sankar, Claire Clivaz, *HumaReC Project: Digital New Testament and Continuous Data Publishing*[17].

SS has been invited to participate to the last meeting of the project Biblia Arabic in Tel-Aviv in November 2017.

CC will deliver a HumaReC paper at the annual SBL 2017 in November 2017 in Boston: "Academic publishing in an Open Access world : a partnership approach".

A video of a talk given for our team, Vital-IT is available on HumaReC here; a teaser can be found on YouTube: https://youtu.be/KerXPKQUJnQ

V. Perspectives

We are periodically monitoring the web activity on humarec.org. Our first analysis of the web traffic over the last 3 months reported a total of 245 sessions, 211 users, a low bounce rate (12%) and approximately 8 min session duration/users. Obviously, our expectations regarding the launch of the platform have been matched. More efforts will now be allocated to consolidating and increasing our basis of periodical users, for example by being more present on scientific social media (e.g.,academia.edu, researchGate) and cross-references by other specialized blogs. The analytical details and methods have also been published as blog article.[18]

It is difficult to evaluate these results properly however, as we do not currently have opportunities for comparison. The frequency and the visibility of the platform are satisfying but interactions with the users are difficult to improve. This aspect of the project invites further evaluation. The implementation of the annotation tool in the manuscript viewer will be a first step toward. If this procedure works well, we could examine the use of similar tools for the rest of the platform.

The transcription and encoding work will be continued in the next months and new folios will be uploaded regularly. SS will continue her research work on the text.

As soon as we have 15,000 transcribed words in Greek and Arabic, we will perform a second test with the Transrkibus tool.

Finally, the main work area will from now be the Web Book: a first public release is planned during August, as the Web book will be part of the Brill catalogue for Biblical studies released this summer.

[17] Among all the new questions we have to discuss in this project, the order of the names of the author is a challenge, because SHS culture and Life Sciences/IT research cultures are different. In this poster, we followed the order usual in the hard sciences: first the person who has prepared the much of content, and at the last position the director of the project. We try to adapt this order depending on the different circumstances. For example, at the annual SBL (see below), only the present people can have their names on the program.

[18] https://humarec.org/index.php/continuous-publications-blog/11-articles/26-aar1.