Digital Editions at the Bibliotheca Hertziana, Rome

Elisa Bastianello

In recent decades, access to bibliographic sources, here intended as manuscripts or printed books, has ceased to be limited to the consultation of the original codices at conservation institutes, or to the publication of printed anastatic or critical editions, because of the increasingly widespread practice of scanning documents. Even at the beginning of the twenty-first century it was not unusual to have to laboriously search to find which libraries had a copy of an incunabulum or a sixteenth-century book, and the work of comparison between certain manuscripts was possible only if the researcher brought with them a print-out (of microfilms) taken from one copy to the libraries that owned the other copies. Now, however, large digitisation campaigns permit this type of study to be carried out from the comfort of home, thanks to colour images with ever-increasing resolution, without limitations in time of the number of documents that might be consulted on any given day.

In addition, the use of standards for the distribution of images oriented towards interoperability, in particular the IIIF,[1] has been gaining ground for some years, thus simplifying the manipulation and reuse of images in online viewers, while optimising access time even to very high-resolution details thanks to the pyramidal zoom, that is, without the need to wait for the entire image to be loaded in order to zoom in.[2]

The ease of access to the images within sources in the historical and artistic field on the one hand requires scholars to carry out a more accurate verification of the contents, which in the past was delegated to previous critical editions precisely because of the difficulty of accessing the originals. On the other hand, these texts are not always suitable for direct quotation, as the final reader might not be able to actually read what is present in the ancient text. Even without entering into the problems of palaeographic decoding of manuscript texts and knowledge of the ancient language, the earliest typography maintained in its first centuries many of the graphic signs, such as ligatures and abbreviations, typical of the manuscript tradition of the time and are difficult to understand for a modern reader who is not specialised. Just think of how characters such as the long 's' (ſ), so similar to 'f', or the use of 'u' also to indicate the consonant sound of 'v' greatly decrease the readability of a printed text. The image of the text to be cited alone is therefore not

[1] International Image Interoperability Framework, https://iiif.io/.

[2] Kelli Babcock, Rachel Di Cresce 'Impact of International Image Interoperability Framework (IIIF) on Digital Repositories' in Kenneth J. Varnum *New Top Technologies Every Librarian Needs to Know: A LITA Guide*, ALA Neal-Schuman, Chicago, 2019, 181-196.

always sufficient to provide all the information to the reader, but imposes the burden of decoding, in the form of transcriptions that expand the abbreviations and normalise the characters.

Among the immediate advantages of transcription, is the possibility of searching for a term directly in the text of the document, without reading it in full or using indexes that have not always been created in a functional way for all types of research. There is also the possibility of using screen readers that automatically read the text aloud (TTS) in cases of reading disability. Among the disadvantages is the enormously time-consuming work required for its implementation. To overcome this disadvantage, there are technologies such as optical character recognition (OCR), which, at least for the most recent printed texts, are able to offer accurate automatic transcription in a very short time. Optical character recognition analyses the characters individually and then uses linguistic vocabularies to identify plausible combinations in the presence of ambiguous characters, such as the letter 'l' and the number '1'. Precisely for this reason, performance with respect to older typography is very low, to the point of making the transcripts unusable. In fact, ligatures, ornamental letters, abbreviations and ambiguous elements make it very difficult to decode texts character by character and the absence of a standardised language further reduces the possibility of disambiguation, producing transcripts that are a jumble of letters and numbers. Yet, despite these great limitations, the access to OCR full text of older books permitted by tools such as Google Books,[3] Internet Archive[4] and many others, has been sufficient in recent years to allow scholars of early texts to do a great deal of work.

Fortunately, in recent years the use of so-called artificial intelligence applied to the recognition of texts, starting with handwritten ones, seems to be able to help improve the situation. Thanks to these technologies it is possible, for example, to take notes on an electronic notebook which are converted into typed text in real time. In the humanities, this has led to the creation of specific projects and platforms for the manual and automatic transcription of handwritten texts, including Transkribus.[5] Transkribus is a platform, currently developed and maintained by the European social cooperative READ-COOP SCE,[6] which allows transcribers, through a desktop app or a web interface, to work, even in a collaborative form, on digital images of texts to create transcriptions, annotating them directly on the corresponding lines of the image. The recognition of text areas, lines of text and the text itself can also be done automatically with the support of HTR (Handwritten text recognition) neural machines that can be trained directly by the user through manual transcriptions, and which are therefore specific to the text

[3] https://books.google.com/.

[4] https://archive.org/

[5] https://transkribus.eu/Transkribus. Philip Kahle, Sebastian Colutto, Günter Hackl and Günter Mühlberger, "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 19-24, doi: 10.1109/ICDAR.2017.307.

[6] https://readcoop.eu/.

and contents. These recognition models can be shared and thus be reused by other scholars who work with documents is somehow correlated, for instance, by type of writing, linguistic content or period. In particular, the neural recognition of texts is based on lines of writing and not on single characters, and can benefit from contextual information, such as how likely it is that two letters can be close together, the way in which they ligate together, thus overcoming the need for a one-to-one relationship between character in the text and in the transcription. This means that in the case of abbreviations, it is possible to automate the expansion of the same as part of the model. Although created for manuscripts, the same neural machines are easily applied to early typography as well, and existing public models can work out of the box.

For this reason, in the spring of 2021 a first collaboration project between the Bibliotheca Hertziana - Max Planck Institute for Art History and READ-COOP was started, with the aim of applying neural recognition to scans of early books, the extant *Rara* Collection, into the institute's digital library DLIB,[7] and in those of two other Max Planck institutes in the humanities, the Kunsthistorisches Institut (KHI) in Florence and the Max Planck Institute for the History of Science (MPIWG) in Berlin.[8] At the time of this writing, the collection contains over 3,800 books, for a total of about 1,300,000 transcribed pages.

Despite having public models available for printed books, in particular Noscemus[9] and Transkribus Print,[10] the main obstacle was to imagine a way to transcribe texts that, while maintaining the practicality of the expansion of abbreviations as implemented in existing public models, allowed the preservation of philological information. The analysis of the use, frequency and type of abbreviations in texts allows the correlation of copies and editions of a text, but makes searching in the text even more difficult. Since it is possible with Transkribus to annotate abbreviations with their expansions as XML tags, a new technology has therefore been developed that can integrate abbreviation tags into the text recognition model, complete with their respective expansions. Although the model is still in its early stages and requires a broader Ground Truth, it has already been applied to texts published before 1520 in the collection, which present a greater frequency in abbreviations.

The books can be publicly consulted online through a special interface, called Read & Search, connected directly to the Transkribus collection, of which it reflects the corrections and additions almost in real time, the pilot version of which is available online from March 2022.[11] The creation of this platform, although based on an existing framework, has required many efforts to allow for the integration of the volume metadata into the search filters. At the moment, the metadata are

[7] https://dlib.biblhertz.it/.

[8] https://dlc.mpg.de/index/.

[9] Stefan Zathammer, 'Noscemus GM 5 [ID 37664]' *Transkribus,* https://www.uibk.ac.at/projects/noscemus/.

[10] READ-COOP, 'Transkribus Print M1 [ID 39995]', *Transkribus.*

[11] https://transkribus.humanitiesconnect.pub

statically connected to the documents of the collection and even if they appear in the search filters, they are not accessible to the reader during the consultation of the text with the transcription. For this reason, a second project is underway for the dynamic connection of bibliographic information through a system of permanent identifiers and a resolver service that guarantees the continuous updating of data in sync with the Kubikat catalogue.[12] In this second project, new models are also being created for the neural recognition of structures, for example the capability of identifying blocks of text relating to the page number, headers, paragraph headings or marginalia and comments. Once these elements are tagged, it will be possible to search for words that are found only in a specific area, such as titles or notes.

The Read & Search platform represents a first form of digital publication, but Transkribus is also suitable for the preparation of transcriptions for edited critical editions. In fact, the transcriptions and annotations made on this platform, which automatically saves in PAGE XML format, can then be exported to other formats such as PDF, docx (Office Open XML) or TEI XML.

A very peculiar use of this tool is the replacement of OCR for the digital reprint of out-of-print books. Although OCR can be very precise when transcribing a twenty-first-century text, it struggles to distinguish font variants such as italics, small caps, superscript, and in general, the reconstruction of the structure requires a lot of manual editing work. The prototype of this genre of digital reprint was John Shearman's book *Raphael in Early Modern Sources – 1483–1602*, published by the Hertziana in 2006 and already out-of-print, being highly sought after. It consists of two large volumes, with a total of over 1,700 pages, containing the transcription of over a thousand period documents with their notes and bibliography. A 'smart' text recognition model has been created, currently not in the public domain because it is under revision, in order not only to recognise the text, but also to add special characters before and after the text variants of each line (for example before and after the superscript number of the reference to the footnote). A layout model was trained as well, to identify the type of text blocks, especially for the sequence of documents, through a different neural machine called P2PaLa.[13] Since the recognition technologies used by Transkribus are only of a visual type, without any Natural Language Processing (i.e., there is no understanding of the content), and the individual pages are analysed individually, it is not possible to automatically recognise if a paragraph is complete in itself or the continuation of a paragraph started on the previous page. This happens for the main text as well as for the notes or bibliographic information. The identification of the text blocks that continue from the previous page is however essential when you want to transform paged text of a paper edition into the continuous text of a digital edition, as the individual paragraphs must be reconstituted. For this reason, a scrupulous human control of all tags was necessary for the correct attribution of the 'continued' tag to the

[12] http://www.kubikat.org/

[13] https://readcoop.eu/transkribus/docu/p2pala/

headless paragraphs. [14] Thanks to this combination of human work and neural machines it was possible to convert and transform the transcription in Transkribus into a TEI XML[15] document, whose first version was published online using the open source TEI Publisher platform.[16] The workflow is a combination of regex substitutions and XSLT transformations that not only join continued paragraph with their beginnings, but also, thanks to the superscript numbers, put all the footnotes in the right position, connected to their reference numbers. Although the project is not yet completed, especially the part concerning the connection between bibliographic references and extended bibliographic records, and the integration of the images, it already allows a structured search within the text of all transcribed documents and comments.[17]

Even if the process of creating the digital reprint lasted more than three months, with costs probably comparable to those required by entrusting the work to an external service, the recognition models and the conversion workflow remain at the Hertziana's disposal for any future similar project. Furthermore, this method grants complete control over the structure and how it is then rendered in the publication phase, which, at the time of the evaluation of offers, was considered very difficult to achieve with traditional tools. All text corrections will automatically go into the Ground Truth of the next version of the model, expanding the benefits for future reuse.

However, the digital publication of a critical edition is not limited to the creation of texts in digital format, but includes critical apparata such as comments, annotations of references and named entities for the generation of dynamic indexes and much more. The pilot project for this genre is the Complete Works of Heinrich Wölfflin, carried out jointly by the Institute of Art History of the University of Zurich and the Department Weddigen of the Bibliotheca Hertziana: so far three of the fourteen planned volumes have been printed. The digital edition, in addition to the critical materials of the new edition, aims to enrich the content with annotations and search functions not possible on the printed page. Given that the project for digital publishing is grafted onto that of a traditional publication already in a very advanced state, it was necessary to create a workflow that took into account the type of materials available. The texts were in fact produced using normal word processors such as MS Word, after having corrected and integrated the OCR of the original volumes. The footnotes of the original editions become automatic footnotes of the text editor, while the editors' comments have been inserted. This, however, creates major problems, since standard word processors do not allow the insertion of endnotes in the footnotes, and therefore these additional comments have been integrated directly into the footnotes in curly brackets. Curly brackets have been

[14] I must acknowledge the support of Viviana Nocerino, Iolanda Pagano and Andrea Pecorella for completing this tedious task.

[15] The conversion workflow was developed with Reto Baumgartner, and will be released as open source once the clean-up and documentation will be finished.

[16] https://teipublisher.com/index.html.

[17] http://rems.humanitiesconnect.pub.

used to indicate numerous elements such as original page numbers, transcription notes, image indications. For this reason, before proceeding with the conversion into TEI XML format, it was necessary to find a way to extract all the comments, transforming even those in curly brackets into footnotes, and differentiate the markers of the other elements through the use of Visual Basic for Applications macros. In addition, a specific MS Word document template was created for the critical edition, containing paragraph and character styles with unique conventional names, referable directly to TEI structural elements. The conversion is then done directly using the Office Open XML with the help of multiple XSLTs. In this case too, the workflow will be available as an open source together with the configuration parameters of the TEI Publisher instance. At the moment, only a demo with volume four of the collection has been published online, as a case study. It is in fact crucial to study the best way to make the content usable, for example by viewing the footnotes directly as pop-ups, comments and editorial information in a side column, or by expanding bibliographic references into complete citations. Alongside the digital conversion of the materials already present in print, an annotation campaign for named entities such as names of people, organisations, places or works has begun, directly on the online edition, thanks to the annotation tool recently made available on TEI Publisher. Thanks to this tool it is possible to easily associate to each entity an identifier in authority files such as GND,[18] GeoNames,[19] Wikidata,[20] or additional customized lists (RDF Local). The annotation tool was also used to signal the presence of errors or omissions with a special tag (TODO). The Complete Works are undergoing a complete re-design with the support of a design company.

The future digital critical edition will hopefully no longer need the Word to TEI conversion phase, since the text might be available as direct TEI export from Transkribus. This mean that a different way for an easy management of editorial commentaries needs to be developed in order to ease the job of editors.

While these projects are not yet finished, they can clearly demonstrate how important it is to work together with existing standards to ensure editions remain available for as long as possible. Because several research funding bodies, especially in the European Union, require Open Access results of funded projects, it is increasingly important to use platforms that can be easily maintained regardless of the project status, long after the funding expires. At the same time, licensing the workflows as open source, will allow researchers to build similar infrastructures without the need of reinventing the wheel from scratch every time, and will enable them to focus more on improvements and content rather than struggling with a new framework.

Elisa Bastianello studied History and Preservation of Environmental and Architectural Heritage at the University of Venice Iuav, where she also obtained her

[18] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.

[19] http://www.geonames.org/

[20] https://www.wikidata.org/.

Ph.D. in History of Architecture and Urban Planning. In 2007 she got a degree in the School of Archival, Palaeographic and Diplomatic Studies at the State Archives of Venice. Between 2013 and 2016, she was part of the scientific team that organized the 500th anniversary of the Ghetto of Venice. Since 2006, she has been part of the editorial board and the webmaster for *La Rivista di Engramma* (online) (http://www.engramma.it/eOS/index.php) at Centro Studi ClassicA Iuav. In 2017 and 2018, she was research fellow at Iuav, where she worked on theatres and celebrations in Italian Courts and on the development of a **semantic platform** (https://burckhardtsource.org) regarding the correspondence of Jacob Burckhardt. Her research ranges from the history of architecture, spaces for music, palaeography in conjunction with digital humanities. In 2019 she joined the Bibliotheca Hertziana as Digital Publications Manager.

elisa.bastianello@biblhertz.it