Navigating the Digital Transition: Lessons from a Hybrid Critical Edition Project

Elisa Bastianello[1], Reto Baumgartner[2]
[1] Bibliotheca Hertziana – Max Planck Institute for Art History, Italy/Germany – elisa.bastianello@biblhertz.it
[2] University of Zurich, Switzerland - reto.baumgartner@zi.uzh.ch

ABSTRACT (ENGLISH)

This article explores the development and publication of a digital critical edition of the complete works of Heinrich Wölfflin, resulting from a collaborative project between the University of Zurich and the Bibliotheca Hertziana – Max Planck Institute for Art History. The project, started as a traditional print edition, aimed to create a digital edition not only of the original volumes but also of the newly published critical volumes, enriched with annotations and facsimile images. The article details the hybrid approach of publishing printed volumes first, followed by their digital versions, and the challenges faced in converting traditional transcription methods into structured digital formats. It highlights the use of legacy tools like Visual Basic for Applications and more modern XSLT transformations to streamline the process. The publication phase emphasizes strategies for ensuring long-term maintenance of the digital edition, including the adoption of open-source platforms and institutional support. The article concludes by reflecting on the collaborative efforts and diverse expertise that contributed to the project's success, offering insights for other teams embarking on similar digital publishing endeavors.
Keywords: Digital Critical Edition; Text Annotation; TEI XML; Open-Source Platforms; Digital Humanities

ABSTRACT (ITALIANO)

Navigare la transizione digitale: lezioni da un progetto di edizione critica ibrida

Questo articolo esplora lo sviluppo e la pubblicazione dell'edizione critica digitale delle opere di Heinrich Wölfflin, risultato di un progetto collaborativo tra l'Università di Zurigo e la Bibliotheca Hertziana – Istituto Max Planck per la storia dell'arte. Il progetto, nato come una edizione critica tradizionale a stampa, mirava a creare un'edizione digitale non tanto dei volumi originali ma di nuovi volumi critici stampati, arricchiti con annotazioni e immagini facsimili. L'articolo descrive l'approccio ibrido nella pubblicazione prima dei volumi stampati, seguiti dalle loro versioni digitali, e le sfide affrontate nel convertire i metodi di trascrizione tradizionali in formati digitali strutturati. Sottolinea l'uso di strumenti del passato come Visual Basic for Applications e più moderni come le trasformazioni XSLT per ottimizzare il processo. La fase di pubblicazione enfatizza le strategie che sono state messe in atto per garantire la manutenzione a lungo termine dell'edizione digitale, inclusa l'adozione di piattaforme open-source e il supporto istituzionale. L'articolo si conclude riflettendo sugli sforzi collaborativi e le diverse competenze che hanno contribuito al successo del progetto, offrendo spunti per altri team che intraprendono iniziative di pubblicazione digitale simili.
Parole chiave: edizione critica digitale; annotazione testuale; TEI XML; piattaforme open source; Digital Humanities

1. INTRODUCTION

In November 2024, following years of meticulous work, the digital critical edition of Heinrich Wölfflin's collected works was formally launched online.[1] This moment presents an opportune occasion to reflect on the journey that culminated in the digital publication of the first two volumes in this series. The project for the complete works of the Swiss art historian represents a collaboration between the University of Zurich and the Bibliotheca Hertziana - Max Planck Institute for Art History in Rome, under the supervision of Tristan Weddigen, Oskar Bätschmann, and Joris van Gastel.

The edition project, initially funded by the Swiss National Science Foundation (SNSF),[2] with additional support by the Bibliotheca Hertziana,[3] evolved beyond its original conception as a traditional printed edition when the first volumes had already reached the typesetting stage. Consequently, the digital edition was developed not merely as a parallel version of the new printed books but rather as a synoptic view of both (and potentially multiple) editions of the original volumes. Beyond the new commentaries and critical

[1] The edition is available at *https://hwgw.humanitiesconnect.pub*.
[2] The SNSF funding began as a research project (grant n. *160081*) from 2015 to 2017 and continued as an editorial project from 2017 to 2024 (grants nn. *157979* and *198243*).
[3] Grant id *BH-P-19-32*, since 2017.

apparatus, the digital edition has been enhanced with an annotation layer that integrates bibliography and named entity linking, alongside IIIF facsimile images of the *editio princeps*. This comprehensive approach ensures that full text search capabilities and index apparatuses are now seamlessly integrated across all volumes.
This article presents the team's experience throughout all stages of creating the digital edition, including challenges and missteps encountered along the way. Our aim is to provide other groups with insights into our process, enabling them to adopt and adapt the aspects that might prove useful for their own digital humanities projects.

2. PREPARATION OF THE EDITION

The first note goes to the fact that this is a hybrid critical edition, where, at least so far, the volumes are first published in traditional printed form and later prepared for the online edition. The edition was grafted onto a long path of traditional transcription with OCR[4] of Heinrich Wölfflin's printed sources that had been manually corrected in the form of MS Word documents. The idea of working with native TEI XML[5] was discarded at the beginning of the project because it was deemed too complex by the team. This choice took into account both the curators' greater familiarity with the editor and the manual workflow used within the publishing house.
The editors, in collaboration with the typesetters, established a list of editorial rules to mark layout and philological information in the text files. These included using double underlining or highlighters to denote different character styles (such as spaced text – Sperrsatz in German), images, or text requiring manual editing during layout. As for the philological notes, including page breaks, location and captions of images, author's pencil notes, and factual errors in the text were indicated by curly brackets. Critical notes were inserted as endnotes, while the author's notes were converted into regular footnotes.
To overcome the limitation of the text editor, which does not allow automatic notes attached to the text of other notes, the critical annotations related to the footnotes of the original text had been supplemented with the use of curly brackets. In many cases, the co-presence of philological information and critical notes led to the nesting of curly brackets. Additionally, while the fonts and sizes of individual elements (e.g., body of the text, footnotes, and citations) had been defined in the editorial rules, they had not been uniformly and correctly applied with character and paragraph styles by the editors, possibly due to the copy-and-paste function that retains the source text style. For these reasons, the files, visually perfect on screen, contained issues such as endnotes occasionally using the headings paragraph style, or titles being just normal text in bold.
The print output, though requiring time-consuming and tedious manual work, was not compromised by the documents' totally incongruent file structure. While not ideal for the printed edition, the challenges were overcome thanks to the fact that it was entirely hand-set layout and it did not create problems in the generation of the typesetting. For instance, notes to the text are in their own InDesign story,[6] disconnected from reference numbers in the text body, that are simple Roman and Arabic numbers formatted as superscript. Even decoding the nested curly brackets was left to the experienced eye of the typesetters.

3. FROM CHAOS TO STRUCTURE

When these same files were used to begin preparing an XML/xHTML version, the total lack of structure in the text become a huge obstacle, making us fear the repetition of the manual typesetting work even for the creation of the base text of the digital edition. At the time, the final version of the InDesign document that was then sent to print was not available, but even when the files arrived, they were not easier to convert than the original unstructured word processor files.
With this in mind, we split the work with one side converting the XML content of the DOCX files into semi-structured TEI, and the other cleaning up the DOCX documents to improve the success chances. The first results were achieved by combining the experience in XML file development and manipulation of the Central IT of the University of Zurich in one hand with long practice in manipulating and normalizing text files delivered by authors in MS Word format for conversion to structured HTML and import in InDesign on the Digital Publications unit of Bibliotheca Hertziana. This collaboration resulted in an initial manual

[4] Optical Character Recognition. On the problems of working with poor quality results (Cordell, 2017).
[5] (*Text Encoding Initiative*, n.d.).
[6] In Adobe InDesign, a story is a group of text frames with flowing text, i.e. main text, appendices, index of names.

prototype of a portion of the text, which was essential for identifying a clean structure template that could be easily transformed between formats.
The built-in style names (i.e. Body Text, Footnote Text, Heading 1) in a template for the word processor (MS Word) are localized between different system and languages. For this reason, it was necessary to create a set of character and paragraph styles fully independent of the program's default styles to ensure consistency among the international research team.
The next step was to adapt the text documents with critical comments to the template. The manual procedure involved identifying local modification of the font (e.g., italicized or underlined text) and applying the correct character style to them. This made the text marked stable when it was necessary to apply the correct paragraph style, cleaning up the text from unwanted styles and returning it to its base character state. Since this was a process repeated several times both on the same file and in different files rather than repeating the sequence of changes by hand, a series of macros were created in Visual Basic for Applications (VBA), then combined together so that they were all executed in the correct order. VBA proved to be the most efficient way to deal with the complex search and replace, based mainly on how the content had been locally formatted.[7]
Once the styles cleanup was finished, it was necessary to distinguish the different uses of curly brackets, converting them, for example, into pseudo tags and font styles. To quickly identify the most common instances, such as page numbers, a regular expression search was used, again within MS Word, to automatically capture and transform as many instances as possible. Since the final plan was to convert the content into TEI XML, when possible, those curled brackets were replaced with "pseudo tags" that could easily be captured and substituted in any subsequent XSLT transformation. The macro for the page numbers, for instance, identifies parenthesis pairs that contain only Arabic or Roman numerals and replaces them with the `<pb>…</pb>` tag while keeping the number originally inside.
The most complex work involved the recognition and extraction of critical notes within the footnotes. Using the TEI note template as a reference, where author's notes are embedded as `<note>…</note>` tags in the superscript reference position, while critical comments are all relegated to a stand-off section of the file, we decided to move the footnotes inline, applying a different color (blue) and enclosing them between a pair of symbols that could not appear in the text (in our case the lower and upper half of the mathematical function symbol). Moving the original footnotes as a part of the main text makes it possible to use the standard endnotes for the critical comment. At this point it was possible to extract the text of the comments in curly brackets (again with a VBA macro), and move their contents to standard automatic endnotes. In order to retain the nested philological annotations, that were planned in curled brackets in the final rendition, it was necessary to temporary capture their content in pseudo tags. This could be performed almost automatically in the most common cases (single words such as {sic!} or references to page numbers) with the support of regular expressions in macros. Unfortunately, in other cases, where the brackets were not properly balanced or there were less common semantic notes or special annotations, hand verification was necessary.
The normalization of the work documents used by the editors of each volume was applied even to work in progress documents, in order to reduce the necessity of inserting curled brackets, although in some cases the editors preferred to wait until the critical work was finished according to previous editorial norms rather than find the notes integrated into the text and insert the comments directly as footnotes. The macros have been organized in a version of the template, which remains separate from that used by the authors because of MS Word's security limitations against documents with macros.[8]
The final text was then converted to TEI XML using the TEI XSLT Stylesheets[9]. Subsequent transformations catch the styles defined by the project-specific template and introduce tags specific to our document types. A series of substitutions based on regular expressions replace the pseudo tags and symbols with the correct XML markup. While the combination of transformations and substitutions brings the danger of breaking the document structure, we found this useful to detect issues in the markup of the MS Word documents which we could fix there and re-try the transformation.
To ease development and to ensure consistency, these steps are orchestrated by scripts specific to the parts of the books.[10]

---

[7] Templates, with and without activated macros, and macros are available as dataset (Bastianello & Baumgartner, 2025).
[8] See note n.7.
[9] (*TEIC/Stylesheets*, 2013/2025).
[10] The scripts and stylesheets are available as dataset (Baumgartner & Bastianello, 2025).

In addition, numerous transformations were applied to create the bibliography and link it correctly to the bibliographic abbreviations from the formatted text.

4. PUBLICATION OF THE TEXT

The second phase is concerned with bringing the text online. Our personal experience with digital editions is that when the publishing interface is created specifically for an edition, there is a high risk that at the end of the funds for the edition the maintenance of the infrastructure is delegated to the goodwill of the team members still working with the institution hosting the edition. This has led to the loss of many editions created since the late 1990s, only in some cases recovered through specific recovery projects in recent years.[11]

Because our digital publishing project depends on funds that, in part, expired on December 2024, several strategies have been developed to ensure long-term maintenance. One is the use of an open-source platform, TEI Publisher, whose maintenance and updating is ensured by an extended community of users and developers independent from the specific edition. The other is to associate the management of the digital edition, once development and coding are completed, with the digital publications manager position at Bibliotheca Hertziana.

At the beginning of development in 2019, the choice fell on TEI Publisher (Turska et al., 2016), at the time in version 6, because the other open-source promising alternative, EVT 2 (Turco et al., 2014, 2019) had not yet been fully developed as a web service. On the one hand, this choice proved successful, because with the inclusion of annotation tools in version 7 (August 2021)[12], linking named entities became intuitive and simple. In fact, it was only a matter of determining the authority file source of the identifiers (in our case the GND[13]) and how to store the information not present in the form of local stored data. Annotations also became a valuable control tool, since the researcher who was in charge of entity tagging could also mark on the fly typos or other items to be rechecked later. At the same time, TEI Publisher's way of interpreting the TEI XML by manipulating the display using ODD[14] is not completely agnostic and requires the insertion of a part of special tags, for example to indicate parallel editions (original edition and new critical printed edition) as milestones. Also, the web components used in the service boxes, for example in the one containing critical comments or images of the facsimile and illustrations, had not yet reached full maturity. This meant a major upgrade of the entire edition, albeit limited to the volume used as a test, during the upgrades to versions 7, 8 and 9. Future version 10, coming soon, is expected to simplify the platform upgrade without requiring updates to the content of the edition, but will require editing of the code prior to the upgrade.

An additional advantage of using TEI Publisher is that it is an application that operates on top of eXist-db, with a powerful indexing system based on Apache Lucene. This makes it possible to configure an integrated structured search of both the individual volume and across the entire critical edition and potentially across several editions. HWGW was created as a stand-alone application and not as a part of a basic instance of TEI Publisher but of a specific application, for which it was possible to change entirely even the appearance of the site that welcomes users.[15] The online edition is fully responsive, to ensure readers can check the content on their desktop and mobile devices.

5. IMAGES AND EXTERNAL SITES

Images in the Digital Critical Edition are available through a IIIF[16] viewer integrated into the TEI Publisher platform. Up to version 8, it was only possible to include individual images using the IIIF image API,[17] while starting with version 9, it is possible to use a manifest JSON file and the presentation API,[18] even if only v.2. This allowed us to integrate additional images such as book covers and blank pages with comments by the author. Furthermore the option to add a local manifest allowed the integration of

[11] A case in point is the "Biblioteca delle fonti storico-artistiche" formerly available at the Scuola Normale Superiore in Pisa, created by Paola Barocchi and restored several times. The relics of the project are available through the Way Back Machine *https://web.archive.org/web/20190913121736/http://fonti-sa.sns.it/fsaInfo.php*.

[12] (*Annotation Editor Released with New TEI Publisher 7.1.0*, n.d.).

[13] Gemeinsame Normdatei (*GND - Homepage*, n.d.).

[14] One Document Does it all (Rahtz & Burnard, 2013).

[15] The instance is available in GitHub (*Biblhertz/Hwgw*, 2025/2025).

[16] International Image Interoperability Framework. (*IIIF Home*, 2025; Kelli & Di Cresce, 2019).

[17] (*Image API 3.0*, 2020).

[18] (*Presentation API 3.0*, 2020).

metadata missing from the official version hosted on the Archive.org servers. The decision to link directly to the source images seemed the most appropriate, as it maintained the direct connection with the Getty Research Institute, which owns the author's personal volumes complete with his manuscript annotations. Unfortunately, in October 2024, a DDoS attack[19] on the archive.org servers forced the institution to shut down most of its public services, and in particular the IIIF server, which had not been restored by the publication of the online edition. Fortunately, the Getty Research Institute allowed the team to publish the images directly from the institutional IIIF server, which is on the same cluster as the critical edition. This experience has made us even more aware of the fragility of using linked resources, even though common and shared protocols such as the IIIF, within a critical edition.

6. FROM THE PROTOTYPE TO THE ONLINE EDITION

Although the first alpha version of the volume edition was developed from the MS Word file sent to the publisher of the print edition, it was necessary to incorporate into the text all the corrections that the editors had added in the layout during proofreading. In particular, moving critical comments into footnotes had revealed that some interlinear annotations, which appeared clear and obvious as long as they were within an in-text citation, became ineffective in footnotes and needed rework and rethinking.
Unfortunately, due to the manual typesetting described above, it was not possible to directly convert the final texts to XML as we did for the MS Word file. It was therefore necessary to export the individual InDesign stories as RTF files and convert them to xHTML. For the volumes where we already had the TEI XML data, we had to compare the text content with the xHTML data. In the future, for volumes that will not be typeset for printing by the publisher, the InDesign typesetting will be generated directly from TEI XML. In this way we can avoid dealing with different versions.
We have developed a comprehensive transformation pipeline that bridges the gap between our TEI XML source files and InDesign layouts[20]. This pipeline uses XSLT transformations to map TEI elements to corresponding InDesign styles through a structured mapping configuration. The process first converts TEI XML to a specialized intermediate format compatible with InDesign's XML import functionality, preserving both structural hierarchy and stylistic information. When importing into InDesign, each element is automatically assigned its corresponding paragraph or character style based on our predefined mapping schema. This bidirectional workflow ensures that any edits made to the source TEI are accurately reflected in the InDesign layout during subsequent updates, while maintaining layout integrity. This approach not only streamlines PDF generation for print outputs but also ensures perfect consistency between the digital TEI XML edition and any print derivatives. By implementing this semi-automated approach, we will significantly simplify maintaining consistency between the TEI XML source and the InDesign layout, eliminating the need for manual synchronization between digital and print versions.
The output is not the only part that will be connected. We have developed a direct transformation between neural OCR transcription in PAGE XML[21] format and TEI XML format for editions of print texts not yet transcribed and for manuscripts.[22] The ideal path was to have editors annotate directly in TEI format, either on TEI Publisher or on an XML editor, but this route was eventually abandoned. Developing a comfortable annotation interface for curators required time and resources that were not available, while annotating directly in XML on code editors required additional technical skills from curators. It was preferred to transform the generated TEI back into MS Word (with inline footnotes) and proceed as usual for the ongoing and future volumes (see Fig. 1).
To ensure easier access to the data, both creation information for the custom TEI Publisher instance [23] and the TEI XML-formatted texts[24] have been published on GitHub. In particular, for long term preservation, the edition data are synchronized on a research data repository that publishes the latest released version of the code, complete with project metadata.[25]

---

[19] Distributed Denial-of-Service (Brewster, 2024)
[20] This workflow is part of the editorial general process at Bibliotheca Hertziana.
[21] Page Analysis and Ground truth Elements (*PRImA*, n.d.).
[22] The code is published as open-source in (*Biblhertz/Trans2tei*, 2021/2024) and was detailed on (Bastianello & Baumgartner, 2023). The workflow uses Pylaia engine with a model trained to add the information and a P2PaLA model for layout tagging, in the Transkribus platform (*Transkribus - Unlocking the Past with AI*, n.d.)
[23] The custom app information is available at (*Biblhertz/Hwgw*, 2025/2025).
[24] The TEI XML data are available at (*Biblhertz/Hwgw-Data*, 2025/2025).
[25] (Bastianello et al., 2025)

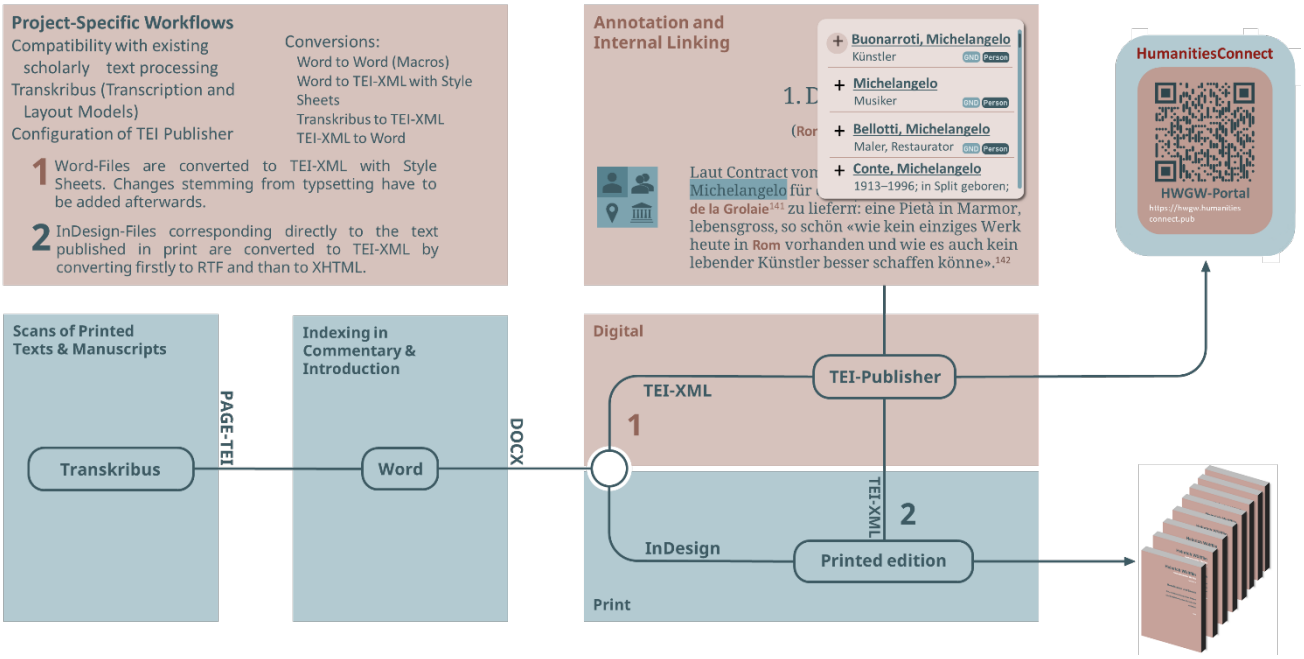

Figure 1 - Schema of the critical edition workflow.

7. CONCLUSIONS

Publishing a digital critical edition from a traditional workflow is not a linear path; it often requires going back and looking at issues from different perspectives. The collaboration of team members with experience and skills in different fields has been one of the key advantages of our working group, so we have decided to make all the development public, in the hope that it will also be of help to other teams, especially those who cannot rely on sufficient funds and such broad expertise.

ACKNOWLEDGEMENTS

This critical edition was funded by SNSF as a research project (grant n. *160081*) from 2015 to 2017 and continued as an editorial project from 2017 to 2024 (grants nn. *157979* and *198243*). Support to the project has been granted by the Bibliotheca Hertziana – Max Planck Institute for Art History since 2017 with grant id *BH-P-19-32*.

REFERENCES

Annotation editor released with new TEI Publisher 7.1.0. (n.d.). Retrieved January 26, 2025, from *https://www.e-editiones.org/posts/annotation-editor-released-with-new-tei-publisher-7-1-0/*.

Bastianello, E., & Baumgartner, R. (2023). L'applicazione del riconoscimento testi neurale per la realizzazione di ristampe digitali. In E. Carbé, G. Lo Piccolo, A. Valenti, & F. Stella (Eds.), *La memoria digitale: Forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD* (pp. 15–23). Associazione per l'Informatica Umanistica e la Cultura Digitale.

Bastianello, E., & Baumgartner, R. (2025). *Heinrich Wölfflins –Gesammelte Werke (HWGW) Digital Edition: Word clean-up conversion scripts* (Version 1.0) [Dataset]. Edmond. *https://doi.org/10.17617/3.4A6MNT*.

Bastianello, E., Baumgartner, R., Meier, S., & Weddigen, T. (2025). *Heinrich Wölfflins –Gesammelte Werke (HWGW) Digital Edition Dataset* (Version 1.0). Edmond. *https://doi.org/10.17617/3.QHJW4D*.

Baumgartner, R., & Bastianello, E. (2025). *Heinrich Wölfflin – Gesammelte Werke (HWGW) Digital Edition: XSLT transformations* (Version 1.0) [Dataset]. Edmond. *https://doi.org/10.17617/3.15WQ42*.

Biblhertz/hwgw. (2025). [HTML]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/hwgw* (Original work published 2025).

Biblhertz/hwgw-data. (2025). [XQuery]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/hwgw-data* (Original work published 2025).

Biblhertz/trans2tei. (2024). [XSLT]. Bibliotheca Hertziana Digital Humanities Lab. *https://github.com/biblhertz/trans2tei* (Original work published 2021).

Brewster, K. (2024, October 18). *Internet Archive Services Update: 2024-10-17 | Internet Archive Blogs*. *https://blog.archive.org/2024/10/18/internet-archive-services-update-2024-10-17/*.

Cordell, R. (2017). "Q i-jtb the Raven": Taking dirty OCR seriously. *Book History*, *20*(1), 188–225. *https://doi.org/10.1353/bh.2017.0006*.

GND - Homepage. (n.d.). Retrieved January 26, 2025, from *https://gnd.network/Webs/gnd/EN/Home/home_node.html*.

Image API 3.0. (2020). *https://iiif.io/api/image/3.0/*.

international image interoperability framework—Home. (2025, January 7). *https://iiif.io/*.

Kelli, B., & Di Cresce, R. (2019). Impact of international image interoperability framework (IIIF) on digital repositories. In K. J. Varnum (Ed.), *New top technologies every librarian needs to know* (pp. 181–196). ALA Neal-Schuman.

Presentation API 3.0. (2020). *https://iiif.io/api/presentation/3.0/*.

PRImA. (n.d.). Retrieved January 26, 2025, from *https://www.primaresearch.org/tools/PAGELibraries*.

Rahtz, S., & Burnard, L. (2013). Reviewing the TEI ODD system. *Proceedings of the 2013 ACM Symposium on Document Engineering*, 193–196. *https://doi.org/10.1145/2494266.2494321*.

TEIC/Stylesheets. (2025). [XSLT]. Text Encoding Initiative Consortium. *https://github.com/TEIC/Stylesheets* (Original work published 2013).

Text Encoding Initiative. (n.d.). Retrieved January 26, 2025, from *https://tei-c.org/*.

Transkribus—Unlocking the past with AI. (n.d.). Retrieved February 15, 2025, from *https://www.transkribus.org/*.

Turco, R. R. D., Buomprisco, G., Pietro, C. D., Kenny, J., Masotti, R., & Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 8*, Article Issue 8. *https://doi.org/10.4000/jtei.1077*.

Turco, R. R. D., Pietro, C. D., & Martignano, C. (2019). Progettazione e implementazione di nuove funzionalità per EVT 2: Lo stato attuale dello sviluppo. *Umanistica Digitale*, *7*, Article 7. *https://doi.org/10.6092/issn.2532-8816/9322*.

Turska, M., Cummings, J., & Rahtz, S. (2016). Challenging the Myth of Presentation in Digital Editions. *Journal of the Text Encoding Initiative*, *Issue 9*, Article Issue 9*. https://doi.org/10.4000/jtei.1453*.