Linked Open Data and IIIF for connecting manuscripts images with their transcriptions: a case study from the Veneranda Biblioteca Ambrosiana

Lorenza Talarico
Catholic University of the Sacred Heart, lorenzatalarico@gmail.com

ABSTRACT (ENGLISH)

This paper presents a case study that leverages Linked Open Data (LOD) and the International Image Interoperability Framework (IIIF) to enhance the accessibility of *De Educandis Ingeniis* by Federico Borromeo, a foundational manuscript of the Veneranda Biblioteca Ambrosiana. This 17th century treatise, reflecting Borromeo's dedication to promoting knowledge for the common good, has been digitized and linked with its textual transcriptions to create an innovative digital edition. The project culminated in a prototype developed on the Simple Annotation Server (SAS), allowing scholars to engage with the manuscript interactively.

The research builds on advancements in digital technologies that have transformed libraries into hubs for cultural dissemination. As the first Italian institution to adopt the IIIF standard, the Ambrosiana offers a rich collection of ancient works through its Digital Library. This study explores how IIIF and LOD can bridge the gap between historical artifacts and contemporary digital platforms, improving usability, accessibility and preservation.

The research was structured into three phases. The first addressed the theoretical foundations of LOD and the Semantic Web while defining the research problem. The second focused on technical implementation, employing Transkribus to create a Handwritten Text Recognition (HTR) model, TEI XML and ALTO formats for transcription, and IIIF standard for image delivery. The final phase developed a prototype on SAS, integrating annotations stored as linked data to ensure interoperability.

This case study demonstrates the integration of cutting-edge technologies to establish a replicable workflow for digital editions of manuscripts and the results highlight improved academic utility and accessibility while bridging tradition and innovation. By proposing a model for linking images and texts, the research addresses the lack of a unified workflow and opens pathways for future development in digital humanities.

Keywords: Linked Open Data; IIIF; Digital library; Transkribus; HTR

ABSTRACT (ITALIANO)

Linked Open Data e IIIF per collegare le immagini dei manoscritti alle loro trascrizioni: un caso studio dalla Veneranda Biblioteca Ambrosiana

Questo articolo presenta un caso di studio che utilizza i Linked Open Data (LOD) e l'International Image Interoperability Framework (IIIF) per migliorare l'accessibilità del *De Educandis Ingeniis* di Federico Borromeo, un manoscritto fondamentale conservato presso la Veneranda Biblioteca Ambrosiana. Questo trattato del XVII secolo, testimonianza dell'impegno di Borromeo nella promozione della conoscenza per il bene comune, è stato digitalizzato e collegato alla sua trascrizione testuale per creare un'edizione digitale innovativa. Il progetto si è concluso con lo sviluppo di un prototipo sul Simple Annotation Server (SAS), che consente agli studiosi di interagire con il manoscritto.

La ricerca si basa sugli sviluppi delle tecnologie digitali che hanno trasformato le biblioteche in centri di diffusione culturale. L'Ambrosiana, prima istituzione italiana ad adottare lo standard IIIF, offre una vasta collezione di opere antiche attraverso la sua biblioteca digitale. Lo studio esplora come IIIF e LOD possano colmare il divario tra artefatti storici e piattaforme digitali contemporanee, migliorando l'usabilità, l'accessibilità e la conservazione dei documenti.

Il lavoro si articola in tre fasi principali. La prima fase ha affrontato le basi teoriche dei LOD e del Web Semantico, definendo il problema di ricerca. La seconda si è concentrata sull'implementazione tecnica, utilizzando Transkribus per creare un modello di riconoscimento della scrittura a mano (HTR), i formati TEI XML e ALTO per la trascrizione e lo standard IIIF per la distribuzione delle immagini. La fase finale ha portato allo sviluppo di un prototipo su SAS, integrando annotazioni archiviate come linked data per garantire l'interoperabilità.

Questo caso studio dimostra come l'integrazione di tecnologie all'avanguardia possa stabilire un workflow replicabile per le edizioni digitali di manoscritti. I risultati evidenziano una maggiore utilità e accessibilità accademica, creando un ponte tra tradizione e innovazione. Proponendo un modello per collegare immagini e testi, la ricerca affronta l'assenza di un workflow unificato e apre la strada a nuovi sviluppi nelle digital humanities.

Parole chiave: Linked Open Data; IIIF; Biblioteca digitale; Transkribus; HTR

1. INTRODUCTION

Nowadays, the integration of digital technologies into historical libraries is proving to be not only inevitable but essential, transforming these institutions from static knowledge repositories into dynamic instruments of cultural dissemination (Cusimano, 2015). This transformation is supported by advancements in technological standards such as Linked Open Data (LOD) and the International Image Interoperability Framework (IIIF), which have enabled the creation of interoperable digital editions accessible to researchers worldwide. Among the few Italian institutions leading this transition is the Veneranda Biblioteca Ambrosiana in Milan. Hence, this case study explores the use of LOD and IIIF standard to link images and text, making one of the Library's manuscripts, *De educandis ingeniis* by Federico Borromeo, accessible in a newly enhanced digital format. That is to say, this brief treatise represents an adequate starting point for testing the integration of digital images with textual transcriptions. Structured into three phases, the approach aims to improve searchability, facilitate academic use and establish a replicable workflow for other manuscripts. In other words, it seeks to answer a central research question: can modern technologies bridge the gap between ancient manuscripts and contemporary digital tools to enhance usability and preservation? The initial phase of the research contextualized the project by addressing the theoretical foundations of LOD and the Semantic Web. The second phase focused on technical implementation, encompassing the use of Transkribus to develop a Handwritten Text Recognition (HTR) model, the employment of TEI XML and ALTO formats and the adoption of IIIF standards. This process culminated in the creation of a final digital edition prototype on the Simple Annotation Server (SAS), enabling the side-by-side visualization of images and their corresponding transcriptions. In fact, the research involved uploading via SAS the annotations of the text transcriptions into the IIIF manifest, sourced directly from the Digital Library of the Ambrosiana, following the principles of Linked Open Data (LOD). As a matter of fact, the LOD paradigm offers significant advantages for language data, such as structural and conceptual interoperability, federation, the expressivity of semantic web models and the network effect. Given the context of this study and its placement within the realm of the Semantic Web, the use of the LOD paradigm was imperative for addressing the research problem. Therefore, the aim was to develop a model that can be replicated and adapted for other manuscripts, enhancing cultural heritage through innovation and accessibility. By bridging traditional practices with advanced technological tools, it addresses the absence of a unified data-structuring model in the community.

2. DATA AND TECHNOLOGIES EMPLOYED

The data at the center of this case study include the IIIF images of *De educandis ingeniis*, published on the Digital Library of the Veneranda Biblioteca Ambrosiana website, along with the corresponding IIIF manifest and the transcription of the manuscript. The Ambrosiana's Digital Library aims to enhance the primary funds, namely Inferior, Superior, S.P. and Trotti, through digitization and open online access. This initiative not only ensures the preservation of these invaluable manuscripts but also promotes their use for study and research. As the first library in Italy to adopt the IIIF standard, the Ambrosiana employs tools like the Mirador viewer and the image server Cantaloupe to provide broad accessibility. Indeed, the IIIF ecosystem is integrated with the library's OPAC (On-line Public Access Catalog), ensuring a direct link between a manuscript's catalog description and its corresponding digital resource (Cusimano, 2019). IIIF, a set of open standards supported by a global consortium of cultural institutions, offers significant advantages, including high-quality image delivery with advanced zooming and manipulation options, compatibility with diverse tools for viewing and comparing images and the flexibility to "publish once and reuse often" across different platforms (Salarelli, 2017).

The transcription of the manuscript was conducted using Transkribus, a platform developed during two EU-funded research projects and maintained by READ-COOP SCE since July 2019. In short, Transkribus, powered by Java technology, facilitates manual and automatic transcription, the training of custom AI models, advanced search tools and exporting documents in various formats. However, Transkribus is a "closed environment", with certain limitations, particularly compatibility issues that arise when attempting to use the work, produced within the platform, externally. These challenges, alongside bureaucratic and authorization obstacles and platform copyright restrictions, collectively hinder seamless adherence to LOD principles. Therefore, the goal of this specific study with *De educandis ingeniis* was to export the work conducted and achieve a similar visualization result outside of the Transkribus environment, to ensure authentic interoperability and alignment with LOD paradigm. For *De Educandis Ingeniis*, a HTR model was created on Transkribus by transcribing the entire manuscript to produce a Gold Standard. The transcription was then exported in multiple formats, including TEI XML, a format widely used in Digital Humanities for encoding and managing digital data. Despite its flexibility and scholarly utility, integrating TEI with IIIF remains challenging and lacks standardization (Monella & Rosselli Del Turco, 2020). As a consequence, after evaluating various approaches, ALTO files exported from Transkribus were then selected for further use.

ALTO, an open XML Schema developed under the EU-funded project METAe, provides a standardized format for describing the layout and content of physical text resources, such as book pages or newspaper articles. The real strength of ALTO format in this case study lies in its ability to serve as a key connecting link, ensuring a smooth workflow without loss of work or redundancy during the annotation phase: it enables efficient transformation, adaptation and linking of data, all within a LOD perspective. In fact, the ALTO files were converted into annotations using the Alto2Annotations transformer and subsequently uploaded to SAS, the last technology employed in this project to create a visualizable prototype via Mirador. SAS, a Java-based annotation server developed by Glen Robson, the Technical Coordinator of the IIIF Consortium, offers an accessible and open-source solution, allowing users to work locally on their computers while sharing the results through links when needed. It supports both IIIF and the Mirador viewer, enabling the creation of annotations stored as linked data in an Apache Jena triple store. This approach aligns with Linked Open Data principles, enhancing interoperability (Ide & Pustejovsky, 2010) and semantic connections across the web and underscores the potential for linked data to serve as a foundation for future and similar projects. However, SAS can experience downtime and this issue arises because this server relies not only on its own stability but also on the stability of the website hosting the referenced IIIF manifest. Since SAS follows the LOD paradigm, everything is interoperable and interconnected, meaning also that if one component fails, it may affect the entire system. Regardless, SAS provides a configurable backend capable of leveraging various technologies, including Jena and Sesame for Resource Description Framework (RDF) support, as well as SOLR. It also supports custom annotation bodies formatted in RDFa, allowing for greater flexibility in how annotations are structured and utilized. Hence, future works could explore these features of the server to fully leverage Linked Open Data and enhance data analysis, enrichment and information retrieval.

3. WORKFLOW

The workflow and methodology of this project started with the study and application of Transkribus within the Veneranda Biblioteca Ambrosiana, but have progressively evolved through research, experimentation and overcoming various challenges. This progression is driven by the fact that the project involves a cutting-edge technology still under development.

The initial step was to select and review the manuscript *De educandis ingeniis* and its history, based on feasibility considerations, such as time available and the manageable amount of data. Subsequently, the manuscript images were correctly uploaded to the Digital Library in IIIF format, alongside the rest of the catalog already available on the platform. *De educandis ingeniis* model was created for the HTR of this type of handwriting, and all 92 pages of the manuscript were consistently and accurately transcribed line by line, forming a Gold Standard. The result, beyond the AI model, was the ability to visualize the manuscript pages, uploaded as JPEG files generated by the "master files", alongside their corresponding transcriptions within the platform, and the work downloadable in various formats. However, it became evident that for data preservation and maintenance, but primarily due to the inability to integrate Transkribus's workflow with the existing IIIF Mirador architecture on the Ambrosiana Digital Library website, further use of Transkribus was no longer necessary. Despite this, the work done on this platform remains relevant, as it provided the HTR model based on the manuscript, serving as the starting point for the entire case study.

Thereafter, exploring the various download formats, an attempt was made to integrate the IIIF images from the Digital Library, specifically the IIIF manifest, with the TEI XML file downloaded from Transkribus containing the transcription. This involved research into previous attempts by the community, particularly the work of Paolo Monella and Fabio Cusimano (Monella & Cusimano, 2019), as well as existing theoretical approaches. In theory, the TEI format meets all interoperability and extensibility requirements (Burnard, 2014), as the IIIF manifest of the manuscript could potentially reference it as an external resource by modifying the manifest with the referenced annotations technique (Monella & Rosselli Del Turco, 2020).

Within the TEI XML file a key element is <*facsimile*>, which connects the transcription to the corresponding images of the manuscript pages. This connection is essential for aligning the digital representation of the manuscript with its physical counterpart, enabling users to view the original image alongside its transcription on compatible platforms, such as EVT (Monella & Rosselli Del Turco, 2020). However, in this case study, EVT, despite being the current state-of-the-art within the community, was not employed. This decision was based on the fact that EVT is primarily tailored for critical editions of philological texts and is designed as a specialized system for a specific use case. As such, it was considered less appropriate for the broader needs of the Digital Library of the Ambrosiana. Consequently, the integration of TEI and IIIF is theoretically possible and experts are actively working on this process. However, for this specific project, no smooth or viable solutions have been found to date.

To address the initial research problem and achieve the project's objectives, an alternative solution was sought to leverage Linked Open Data for displaying manuscript images alongside their transcriptions. This research led to the discovery of SAS, which aligns with the LOD paradigm by storing annotations as linked data in an Apache Jena triple store. Compatible with IIIF and the Mirador viewer, SAS enhances interoperability and semantic connections across the web, reflecting LOD principles and maximizing the potential for data integration. Additionally, this server facilitates the annotation of IIIF manifests and enables their automatic modification while allowing the upload of JSON files as annotations. This ensures the preservation of previous work and enables a smooth workflow without unnecessary delays. Regarding *De educandis ingeniis*, 92 XML files, each corresponding to a manuscript page, were indeed exported from Transkribus in ALTO format, chosen for its compatibility with SAS. ALTO is widely utilized in OCR and its integration with the Alto2annotations transformer facilitated the conversion of these files into JSON annotations that are compliant with IIIF standards. This process ensured the seamless transfer of work from Transkribus to SAS, effectively linking the transcriptions to each manuscript page available online and maintaining an efficient workflow throughout the annotation process.

The objective was to establish a single prototype for this project utilizing the modified IIIF manifest through SAS: this, thus, involved structuring the data via the latter and revising the IIIF manifest obtained through a link from the Digital Library. The modification employs the other content technique, which incorporates links to Annotation Lists that point to external resources, namely, JSON files created by SAS containing the annotations written for each manuscript page. Before achieving this, it was necessary to convert all 92 ALTO files using the Alto2annotations transformer. However, the first visualization in Mirador showed that the rectangles did not correspond to the lines of text, resulting in disorganized transcriptions. To address this issue, line level annotations were utilized (instead of word level ones) and the coordinates were removed from each file, allowing the annotation for each page to refer to the entire corresponding page. SAS required JSON files compatible with Open Annotations IIIF v.2, necessitating the use of the "annotationListNoArt.xsl" XSLT from the transformer. Subsequently, the output JSON files were manually modified to include the URI of the corresponding canvas from the published IIIF manifest in the Digital Library, ensuring proper communication and interoperability. Once the files were prepared, all annotations were successfully uploaded to SAS. Hence, the server generated the modified IIIF manifest and allows for visualization in Mirador or Universal Viewer, as well as sharing the work via a link.

As mentioned before, SAS employed Annotation Lists to add annotations, specifically transcriptions, as this technique aligns with IIIF version 2. Annotation Lists are a method of grouping annotations, often at the canvas or page level. They may include resources like page transcriptions or OCR data formatted as annotations. These lists are linked to the specific canvas they pertain to, serving as independent resources that should be accessed separately when needed. This separation from the main manifest allows clients to display images quickly, enhancing user experience by adding additional content and commentary as the user interacts with individual canvases. Additionally, Annotation Lists are typically resolvable, meaning their *@id* can be entered in a web browser to retrieve the list directly. As a result, they can be generated dynamically, while the manifest remains static and easily cached. In the URI pattern, the *{name}* parameter should be unique to distinguish one Annotation List from others. This parameter often matches the canvas name, though it is not mandatory, as a single canvas might contain multiple lists categorized by resource type. Therefore, it is important to follow the specific URIs provided, rather than generating them in advance. Furthermore, each Annotation List must have a unique HTTP(S) URI specified in the *@id* field, which should return a JSON representation when accessed. Annotations are included in a *resources list* and the format should match the media type returned upon dereferencing the resource. The *content* type should be selected from a list in the Open Annotation specification or other widely recognized ontologies. For resources directly rendered, like images, transcriptions or musical performances, the motivation should be *sc:painting*. Content resources can also contain additional fields, such as labels, descriptions, metadata, licenses and attribution.

In the context of this case study, the JSON file of the IIIF manifest published in the Digital Library includes the key elements which are characteristic of a manifest. Within the *sequences* property, the canvases are listed. Upon examining the modified and annotated IIIF manifest generated through SAS, it appears very similar to the original one. However, when analyzing the *sequences*, a new property named *otherContent* is found following the *thumbnail* property: it has been utilized by SAS to link additional content, specifically an Annotation List, to the manifest. This represents the main part of the new manifest, with the *@id* linking to another JSON file, specifically an Annotation List for each page of the manuscript. It is important to note that this structure is consistent across all 92 images, and thus applies to every canvas within the manifest in the same way. In analyzing the JSON of the Annotation List, the text of the page transcription in the text/html format and the motivation of the annotation, specifically *sc:painting*, are immediately evident and this confirms the theoretical framework discussed previously. With this structured, interoperable, precise and hierarchical data organization, the desired and final prototype becomes accessible on Mirador.

4. RESULTS AND DISCUSSION

In Mirador the manuscript can be virtually browsed page by page, with each image featuring an annotation corresponding to the text transcription completed at the project's outset using Transkribus. This approach successfully achieves the initial objective of the research: to add the transcription alongside the manuscript and link it to the IIIF manifest of the manuscript's images. This enhances interoperability and provides the Ambrosiana Library with a methodology to potentially further develop in the future its digital library system. The goal is thus to make its invaluable collection even more accessible, interoperable and reusable for researchers, students and enthusiasts worldwide (Cusimano, 2021).

Hence, this work proposes a solution for integrating images and text within the Ambrosiana Digital Library, employing LOD and IIIF technologies to make the manuscript *De educandis ingeniis* accessible in a newly digital and interactive format. The workflow was designed to overcome various bureaucratic, technical and copyright limitations associated with Transkribus, as well as the current lack of necessary Mirador plugins in the Ambrosiana Digital Library for annotating IIIF images. By progressing from transcription in Transkribus to annotation in SAS through multiple steps of format conversion, a digital edition prototype was developed, offering both accessibility and interactive research opportunities to scholars and enthusiasts. Although the results remain within the research domain and the prototype is not yet available on the Ambrosiana website, they highlight how LOD and IIIF could significantly facilitate the navigation, search and consultation of digitized manuscripts while preserving the quality and integrity of the original materials. This model also shows promise for replication with other manuscripts and even across other libraries, thus supporting a broader and more systematic dissemination of cultural heritage (Spina, 2023).

Nevertheless, certain constraints were encountered, as these technologies represent cutting-edge advancements that continue to evolve. Restrictions within Transkribus's closed ecosystem and the challenges of managing complex data partially limited the workflow's effectiveness in a standardized digitization context. As a consequence, the community shall prioritize research on the link between text and images and future developments could include collaborations with other platforms and institutions to enhance interoperability and accessibility, as well as establishing a unified and shareable data structuring model. However, future works should also encompass other technical improvements, such as supporting alternative encoding strategies to enable links to external resources, while ensuring that the entire edition, including XML, internal and external data and software, remains sustainable, durable and aligned with the FAIR principles (Wilkinson, Dumontier, Aalbersberg et al, 2016). Moreover, continuing to use the SAS platform, it would be beneficial to conduct an in-depth study on the coordinates required for accurate annotations. It would be valuable to establish and standardize a method to convert image coordinates from ALTO format via Transkribus to IIIF format. That is to say, while this essential conversion is technically feasible, its implementation remains challenging. If successfully completed, it would help meet the parameters required by the transformer on SAS, enabling word level annotation. This, in turn, would lead to a more sophisticated, precise and efficient prototype that closely resembles the output of Transkribus. Furthermore, it is clear that the future aim would be to apply this workflow to an increasing number of manuscripts, initially from the Ambrosiana Library and later from other institutions, continually refining the process to make it more streamlined and efficient: integrating manuscript images with a searchable text edition is a critical advancement for the academic world, as there currently appears to be no established standard for achieving this goal (Spina, 2023). In this context, the adoption of LOD is not merely conceptual, but functional: the project embraces structural and conceptual interoperability, the expressiveness of RDF-based models and the potential for data federation. In future developments, data enrichment could be enhanced, cross-referencing with external authority datasets and improve information retrieval capabilities across the cultural heritage domain.

In conclusion, this project currently represents a small yet valuable contribution to the community and cultural resource management, fostering an ongoing dialogue between innovation and tradition. The developed prototype can facilitate the digital preservation of manuscripts while enhancing their accessibility and usability, echoing the ideal of *ad publicum commodum et utilitatem*, namely the common good, which inspired Federico Borromeo in founding the Veneranda Biblioteca Ambrosiana. Ultimately, history repeats itself: while contexts and technologies may evolve and change, the enduring spirit of public service and cultural sharing remains a constant in society.

REFERENCES

Burnard, L. (2014). What Is the Text Encoding Initiative?, OpenEdition Press, Marseille.

Cusimano, F. (2015). Bibliotecario digitale, umanista informatico. Bibelot. Notizie dalle biblioteche toscane, 21, 14-17.

Cusimano, F. (2019). La nuova biblioteca digitale della Veneranda Biblioteca Ambrosiana. L'Almanacco Bibliografico. Bollettino trimestrale di informazione sulla storia del libro e delle biblioteche in Italia, La questione, Edizioni CUSL – Università Cattolica del Sacro Cuore, Milano, 1-2.

Cusimano, F. (2021). The IIIF-based Digital Library of the Veneranda Biblioteca Ambrosiana. Umanistica Digitale, 10, 423-432.

Ide, N. & Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. Proceedings of the Second International Conference on Global Interoperability for Language Resources, Hong Kong, China.

Monella, P. & Cusimano, F. (2019). Linking Text and image: TEI XML and IIIF, http://www.paolomonella.it/reires2019/index.html.

Monella, P. & Rosselli Del Turco, R. (2020). Extending the DSE: LOD Support and TEI/IIIF Integration in EVT. Marras, C., Passarotti, M.C., Franzini, G. & Litta, E. (2020). Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica, Università Cattolica del Sacro Cuore, AIUCD, Milano.

Salarelli, A. (2017). International Image Interoperability Framework (IIIF): una panoramica. JLIS.it, 8, 50-66.

Spina, S. (2023). Handwritten Text Recognition as a digital perspective of Archival Science. AIDAinformazioni, 115-132.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18