Digital Humanities Intersections (DHI)

Introducing the TEI (Text Encoding Initiative) Framework for Minimal Digital Scholarly Editions

Saniya Irfan

Department of Humanities and Social Sciences, Indian Institute of Technology, Delhi, India

Research Article

Cite this article: Saniya Irfan. 2026. "Introducing the TEI (Text Encoding Initiative) Framework for Minimal Digital Scholarly Editions" Digital Humanities Intersections, DOI: 10.66244/105392071

Keywords:
Text Encoding Initiative (TEI), Scholarly Digital Edition (SDE), Indian DH, Humanities Data, Encoding

Corresponding Author:
Saniya Irfan
Email:
huz228239@iitd.ac.in

Abstract

This study presents a framework for encoding data in the humanities and social sciences according to the Text Encoding Initiative (TEI) guidelines. The notion of TEI is not widely recognised in the Indian digital humanities (DH) community. There remains a wide variety of South Asian literature that can utilise TEI guidelines to enhance texts and facilitate their use by the academic community both within India and beyond, catering to novices and specialised scholars for broader accessibility. TEI is utilised by global communities in various languages for the production of digital editions, serving diverse audiences. Some editions are enriched with scholarly information, while others function merely as online archives, facilitating further textual processing. This method paper presents an organised structure for the creation of a minimal 'digital scholarly edition (DSE)', a concept defined and borrowed from Sahle (2016) as an archive of significant academic work. The study elucidates the process from data extraction to preprocessing, followed by a detailed guide on encoding data with TEI tags, highlighting the significance of markup and the role of TEI in digital archiving, particularly for humanities data. The digitisation of cultural manuscripts is essential as it aids the preservation of original documents, increases accessibility, and reduces the necessity for direct handling of frequently consulted rare manuscripts. In the forthcoming era, Indian DH must adopt innovative techniques to encapsulate material with embedded metadata for enhanced preservation strategies. A DSE is a critical depiction of historical materials, carefully curated and digitally displayed to enhance accessibility and understanding for scholars and the general public. Through the use of digital tools, DSEs can offer various text versions, emphasise differences and deliver comprehensive analyses that are impractical in print formats. This method maintains the original content while augmenting its applicability and significance in modern academia. The TEI is a methodology for converting unstructured, plain digitised text into a digital scholarly edition (DSE) that incorporates encoded metadata, thereby enhancing information retrieval, computational analysis and visual representation. The term 'minimal' denotes the fundamental initial process of encoding documents according to TEI guidelines. Given that encoding can be a tedious process depending upon the specific tagging required for producing a digital edition tailored to a particular audience, this paper restricts itself to presenting only the basic encoding principles at an introductory level. This aims to familiarise readers with TEI and promote the digitisation of the extensive literary heritage of South Asia for broader engagement.

Introduction

Interacting with an eighteenth-century manuscript written in traditional dialects requires attention to the common formats, which are usually available in hard

copy or as scanned images. For humanities scholars, the dependence exclusively on scanned images has constraints, rendering them inadequate as primary materials for rigorous research activities. The scanned images of texts are not machine-readable; hence, we are unable to search, index or conduct text analysis. They also restrict accessibility for visually impaired individuals, and many digital tools cannot utilise images. Moreover, there is an absence of semantic structure, preventing the ability to annotate names, locations, verses or variants, which could otherwise facilitate computational literary studies. The reusability factor is inadequate, necessitating the repetition of transcription and optical character recognition (OCR) processes for each new project.

To tackle this problem, the Text Encoding Initiative (TEI) provides a standardised framework to preserve and ensure access to historically and academically significant literature. This framework complies with global standards, ensuring the preservation of the content with added layers of semantic or structural information, thereby facilitating accessibility, reusability and long-term archival storage. It offers organised, machine-readable text-encoded files and facilitates semantic markup for entities such as individuals, locations, events and verse forms. Additionally, it facilitates searching, connecting and analysing across huge corpora, guarantees interoperability and reusability across many platforms, and supports diverse research enquiries in future projects, which can leverage existing data while incorporating additional layers of information. It also improves the representation and documentation of the original source while maintaining its citable status.

Cultural heritage represents a valuable inheritance that transcends international boundaries and should be preserved for the benefit of future generations (Singh 2012). The digitisation of cultural manuscripts is crucial since it improves the preservation of original documents, enhances accessibility and minimises the need to physically handle extensively used original manuscripts, particularly those that are rare (Bansode 2008). These literary archives have historically benefited scholars by providing free online access to previously inaccessible collections. However, most of this rare material is on the verge of extinction due to the deterioration of the ageing analogue media. There is an urgent need to preserve this material by converting it to digital formats to save the loss (Jaswal 2016). Digital scans of textual literary materials have facilitated qualitative research enquiries; however, such scans remain inadequate for conducting further quantitative analysis of the texts. Text Encoding Initiative serves as a methodology for the in-depth examination of humanities and social sciences texts using a human-in-the-loop framework in the digital analysis of texts. This approach preserves cultural heritage while providing researchers an opportunity to expand upon their original qualitative research enquiries and integrate additional analyses from academia.

The archival process of preserving books and manuscripts in Urdu literary circles is a simplistic procedure, yielding only scanned reproductions of printed or handwritten texts. The scanned photos are un-editable PDF files that are non-machine readable (Taj and Gala 2024; Ijaz et al. 2025). Rekhta[1], a prominent online Urdu archive in South Asia, along with Anjuman Taraqqi Urdu[2], the oldest institution dedicated to the preservation and promotion of the Urdu language in India, and multiple other platforms, provide only static images of literary materials for scholars and the public to utilise. Although digital archives are crucial for preservation, they frequently provide texts in a 'flat' format, i.e., either as images or simple transcriptions, rendering the words accessible while obscuring the material and historical contexts of the text. A photocopy of a manuscript retains its content while obscuring the indicators of authorship, including deletions, insertions and marginalia. Conversely, TEI facilitates a stratified representation of texts; the original transcription is maintained, while supplementary layers of code document editorial modifications, physical attributes and interpretive components. This converts the text from a static artefact into a dynamic scholarly entity that may be interpreted on numerous levels concurrently. TEI serves not just as a digitisation tool but also as a platform that bridges the physical and mental dimensions of a text. The need for TEI-based scholarly editions stems from their capacity to preserve the integrity of the original while supplementing it with structured annotation, so highlighting the subtleties of textual materiality that basic digital archives fail to represent.

Machine-readable digital artifacts can bring about several benefits such as (i) *Enhanced Accessibility*—users can readily search and locate specific information if the text is machine readable[3], (ii) *Interactive Features*—users can engage with the text through hyperlinks, multimedia integrations and interaction glossaries (Clinton-Lisell et al. 2023), (iii) *Contextual Annotations*—annotations can provide explanations, translations, or cultural context, aiding deeper comprehension and interest in the text and thereby in the language (Arko et al. 2006), and (iv) *Preservation of Linguistic Data*—machine-readable texts can be analysed computationally, thereby supporting linguistic research and development of language learning tools. For instance, at the Schiller-National

[1] https://rekhta.org/

[2] https://atuh.org

Please refer to the following links for more info on interactive Digital Archives:

[3] https://guides.loc.gov/chronicling-america/improved-text

Museum[4] in Marbach, Germany, visitors can read historical manuscripts in German, some of which are difficult to decipher. Utilising provided digital devices, they can access translations and contextual edits by positioning the device over the original manuscripts, thereby gaining insight into the preserved material for both visitors and scholars.

Procedures that necessitate text in a digital format for achieving these benefits and text analysis using computational models and software cannot utilise scanned, non-machine-readable and un-editable texts, rendering them unsuitable for digital examination. A multitude of rare and classic text collections in South Asian languages are now accessible in online archives following extensive digitisation and preservation efforts (Taj and Gala 2024; Jaswal 2016); nonetheless, there remains an urgent need to convert them into layered digital scholarly editions (DSE). Such editions are critical representation of historical documents, meticulously prepared and presented digitally to make historical texts more accessible and comprehensible to scholars and the public (Sahle 2016). Digital editions possess distinct features that become immediately apparent. While print versions can be turned into electronic texts and digital publications to make use of some of those features—these include discoverability (being able to find), usability (being able to use) and computational access (being able to compute), digital editions go beyond these advantages. Digital versions have other, more important features that come from changes in how they are prepared, the methods used and the theories that support them. You could say that digital editions follow a digital paradigm, just like printed versions have been following a paradigm that was shaped by the technical limits and cultural norms of typography and book printing. It's not possible to fully understand what a truly digital model means just by digitising printed materials. To put it simply, 'A digitised edition is not a digital edition' (Sahle 2016).

For Sahle, the terms 'scholarly digital edition' and 'digital scholarly edition' can be used interchangeably. A scholarly digital edition (SDE) would underscore the phenomena of digital publication and the necessity of ensuring its scholarly quality at this juncture. This would include incorporating a critical dimension into otherwise potentially non-critical articles. While, the digital scholarly edition pertains to the traditions and practices of scholarly editions, illustrating their evolution into the digital domain. As both yield identical outcomes, they do not inherently signify distinct concepts; thus, aside from an in-depth discourse on methodologies and interpretations, they may be regarded as synonyms (Sahle 2016). accordingly, Sahle defines DSEs or SDEs as 'scholarly editions that are guided by a digital paradigm in their theory, method and practice' (Sahle 2016).

By utilising digital tools, DSEs can present multiple text versions, highlight variations and provide in-depth analyses, that are not feasible in print editions. This approach preserves the original material as well as enhances its usability and relevance in contemporary scholarship (Sahle 2016). The TEI is a method for converting flat—i.e., having no structure, unlike a structure in TEI—plain digitised text into a DSE containing encoded metadata, which facilitates enhanced information retrieval, computational analysis, as well as a visual representation.

Literature Review

The Text Encoding Initiative Consortium, or TEI Consortium, is an international consortium that is dedicated to maintaining TEI guidelines as a recommended standard for textual markup[5]. The TEI grew out of a recognised need for the creation of international standards for textual markup that resulted in a conference at Vassar College, Poughkeepsie, in November 1987. The overall purpose of the project was to produce a set of guidelines for the creation and use of electronic texts in the majority of linguistic and literary disciplines (Cummings 2013). Given below are a few reasons for the encoding of texts in an electronic format:

- To make explicit (to a machine) what is implicit (to a human)
- To add value by adding annotation and tags
- To facilitate reuse of the material across different formats, contexts and users

Humanities scholars do not very often compute; but many of them do categorise and analyse, and those are aspects that they can appreciate. More profoundly, perhaps, they do all communicate. Marking up a text involves both analysing it and communicating that analysis (Roueché 2012).

Scholarly Digital Edition using Text Encoding Initiative Guidelines

Patrick Sahle defines a DSE as:

> A scholarly edition is the critical representation of historical documents that often stand for a certain text or work. Here, we may talk about accessibility, searchability, usability, and computability. However, a main characteristic of a digital edition is its representation of a potentially large number of documents in a potentially limitless number of different views, such as facsimile, diplomatic transcription, and reading versions. All are generated from the same electronic code according to certain, sometimes even user-controlled, modulations (Sahle 2016).

[4] https://www.dla-marbach.de/museen/schiller-nationalmuseum/

[5] https://tei-c.org

Text Encoding Initiative or TEI is a set of standard guidelines for encoding humanities texts using XML, or Extensible Markup Language. It is also used by specialised research projects using more detailed markup to represent features of the text that are relevant to specific objectives of the research, such as linguistic analysis, or to serve as the basis for a dictionary. Encoding fundamentally refers to the transformation of text into a machine-readable format. We generally incorporate XML tags into the unrefined Natural Language content, providing a structured framework to the otherwise unstructured raw material. Within this framework, TEI tags impart meaning to the structured material within XML. Once encoded, the structured text can be queried to extract information based on any criteria, thus facilitating data analysis. A multitude of TEI tags exist to encode each type of characteristic.

There are several humanities projects that use the TEI interface, such as the following:

1. **The Women Writer's Project (WWP) by Northeastern University**[6] is one of the foremost encoding projects using TEI guidelines to study various objectives like intertextuality and reception history in the selected corpus of 1500–1850.
2. **Search and Retrieval of Indic Texts (SARIT)**[7] is a collection of texts in Sanskrit maintained by The British Association for South Asian Studies and the Institute for the Cultural and Intellectual History of Asia at the Austrian Academy of Sciences in Vienna.
3. **The Decameron Web by Brown University Scholarly Technology Group**[8] is a project designed to study the Italian text through flexible, well-structured digital resources on the literary, historical, and cultural context of *Decameron* by Giovanni Boccaccio.

Tools Utilised

This methodological paper employs the following tools:

1. **OCR Tool:** We utilised the widely renowned online open-source platform ILovePDF[9] for the extraction of unrefined text from the scanned pages. Our study is conducted on an English translation, and numerous open-source OCR technologies are proficient with English texts; consequently, utilising premium choices such as Google OCR is advisable for scripts where standard OCR engines underperform. Consequently, we restrict our usage to ILovePDF instead of alternative Python libraries such as Tesseract or Google OCR, which necessitate the user have supplementary programming expertise.
2. **R Programming:** The photographs were scanned in a landscape orientation, complicating their processing with the OCR tool, as the identified text was non-linear, while the program is optimised for portrait images. To resolve this issue, we utilised R code[10] to treat the two scanned pages as a single PDF page, instead of manually trimming them. The code utilised pixels and page measures to obtain a singular scanned image collection.
3. **oXygen XML Editor:** oXygen XML Editor[11] is the premier tool for XML authoring and development. Designed for all users, ranging from novices to specialists, it is adaptable, cross-platform and accessible as both a standalone application and a plugin. Featuring strong support for XML technologies, it provides tools for efficient content generation, editing and publishing. Utilising XML templates from the oXygen XML Editor presents numerous specific benefits, particularly when the objective is to produce a TEI edition. In contrast to standard XML editors or web platforms, oXygen offers a TEI-integrated environment that facilitates both structural and semantic encoding procedures.
   First, oXygen provides pre-existing TEI templates (including TEI All, TEI Lite and TEI Corpus), which are equipped with preconfigured schemas, namespace declarations and validation criteria. This guarantees that each tag utilised adheres to TEI standards, minimising human mistake and maintaining sustained compatibility. Additionally, it offers a sophisticated editing interface featuring auto-completion, tooltips and documentation for TEI elements and attributes, facilitating the selection of appropriate tags for encoding individuals, locations, textual divisions or editorial annotations. Primarily, oXygen streamlines the formulation of the TEI header, which is frequently intricate yet essential for academic editions. It offers organised prompts for components such as '<teiHeader>', '<fileDesc>', '<sourceDesc>' and '<encodingDesc>', directing the user to document metadata like title, author, publication history, transcription methodology and editorial guidelines. This guarantees that your TEI file is not just acceptable XML but also a thoroughly documented digital academic artefact.
   The oXygen XML Editor facilitates TEI work by enabling scholars to encode, validate, visualise and manage TEI documents on a single platform, providing enhanced metadata control, automatic validation and meticulous tag management in contrast to generic XML or online tools.

[6] https://wwp.northeastern.edu
[7] http://sarit.indology.info
[8] http://www.brown.edu/Departments/Italian_Studies/dweb/index.php
[9] https://www.ilovepdf.com
[10] https://www.r-project.org
[11] https://www.oxygenxml.com

4. **TEIGarage:** We utilised OXGarage (now referred to as TEIGarage)[12], a web-based platform enabling users to convert files among diverse scholarly text formats, including TEI XML and XHTML (Extensible Hypertext Markup Language). The XML file is converted to XHTML by uploading the XML file on the OXGarage interface, selecting 'XML (TEI)' as the input format and 'XHTML' as the output format, and then executing the conversion. The system autonomously implements TEI stylesheets to convert structured XML data into a web-compatible XHTML format accessible in any browser.
   XML is a versatile markup framework intended for the storage and transmission of structured data; it establishes custom tags that delineate the semantics and structure of information. XHTML (Extensible Hypertext Markup Language) is a particular use of XML that adheres to more stringent syntax requirements for representing web pages. XML emphasises data representation and interchange, whereas XHTML is designed for web display, serving as a well-formed, XML-compliant variant of HTML that allows web browsers to produce structured content with more reliability.
5. **Sublime Text:** Sublime Text[13] is a simple text editor enabling academics to examine and modify the foundational structure of digital texts, including XML and XHTML files. It is frequently utilised in digital humanities projects due to its provision of a clear, legible environment for markup manipulation without the necessity of extensive programming expertise. In the development of a TEI-based digital edition, Cascading Style Sheets (CSS) are frequently employed to manage the visual presentation of the encoded text on the display. This indicates that the TEI or XHTML file encompasses the structured material—such as names, locations or portions of a poem—whereas the CSS file dictates the visual presentation of these elements in a web browser. For instance, personal names may be shown in italics or events in an alternate hue. Connecting the CSS file to the XHTML version of the TEI text enhances the digital edition's visual appeal and readability, while preserving the foundational scholarly markup. Stylesheets are readily accessible online; therefore, we utilised those from W3Schools[14] and incorporated them into the XHTML file of our digital version.

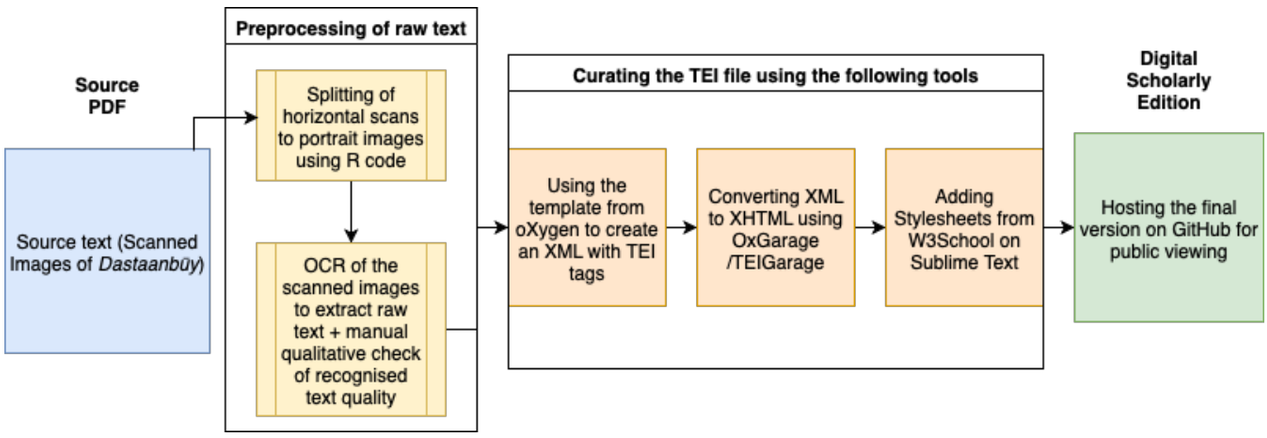


Figure 1: High-level pipeline used in the paper

The TEI Schema

Apparent in its nomenclature, text markup entails the delineation of distinct segments within a text through the utilisation of numerous elements incorporated within the TEI schema. The TEI employs a hierarchical, branched structure to encapsulate textual content. This tree-like configuration originates from the primary root element, denoted as <tei>, and necessitates two obligatory sub-elements: <tei header> and <text>.

The presented image (Figure 2) originates from an XML file encoded through the application of a predefined template within the Oxygen XML editor[15], encompassing an array of available elements. Users have the flexibility to select templates that align with their specific requirements for incorporating elements into their textual content. The hierarchical structure of this encoding adheres to parent, grandparent and sibling relationships, delineating that a sibling element cannot exist at the level of a parent element. Following the two compulsory sub-elements previously delineated (<tei header> and <text>), the subsequent element in the hierarchy is the element. The selection of this element is contingent upon the organisational structure inherent in the text to be encoded, whether it be a book, chapter, essay, poem, act, and so forth.

[12] https://teigarage.tei-c.org
[13] https://www.sublimetext.com
[14] https://www.w3schools.com/css/default.asp
[15] https://www.oxygenxml.com

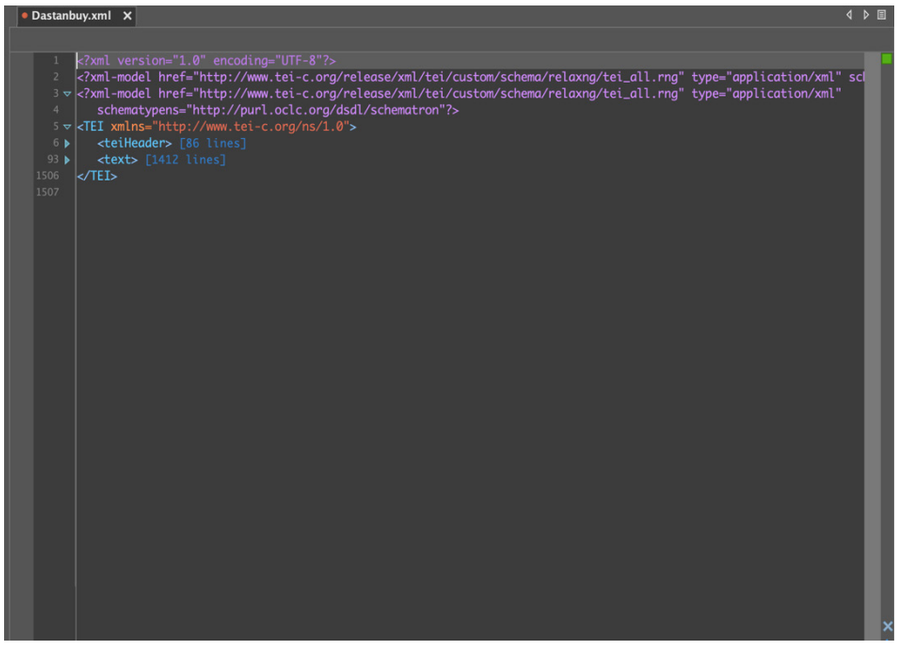


Figure 2: Root elements

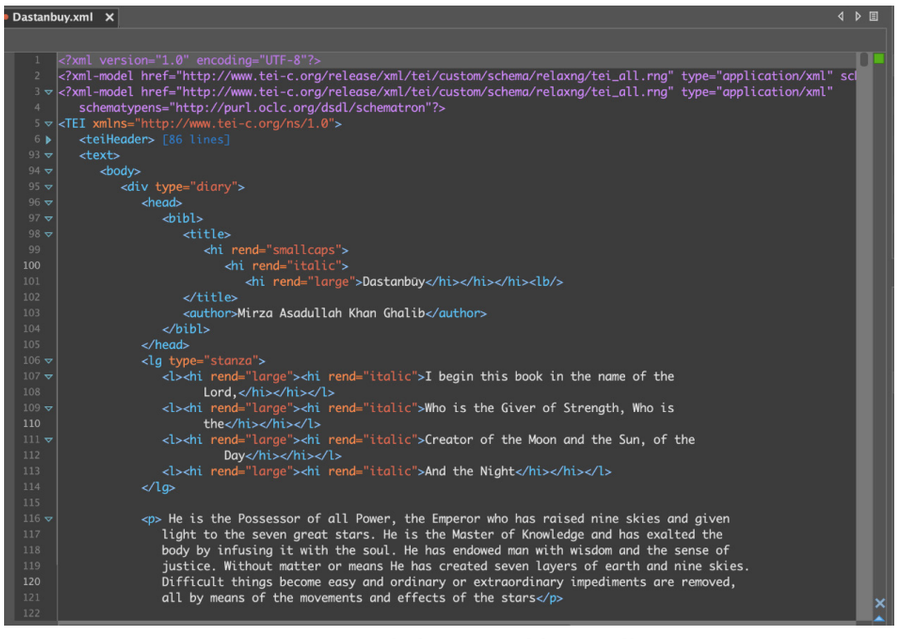


Figure 3: Branched structure of the XML file

The adoption of XML within the TEI is underpinned by several advantageous attributes. XML is human-readable and comprehensible, facilitating ease of interpretation. Its coding intricacy is relatively modest, rendering it accessible for users. Furthermore, XML is entirely portable, ensuring compatibility with diverse platforms. As a preferred archival format, XML is endorsed for long-term preservation initiatives.

Note that the hierarchical structure of XML encoded text is apparently quite visible in the manner of positioning of the elements in the above screenshot (Figure 3) from oXygen XML editor.

Many documents and media have been entered in databases that hold content in tables, records and fields exportable into XML. An increasing number of businesses, publishers, booksellers, university libraries and digital text archives now use databases and XML to manage the jostling, dynamic bundle of data objects we once called books, articles, reports or songs (Liu 2004).

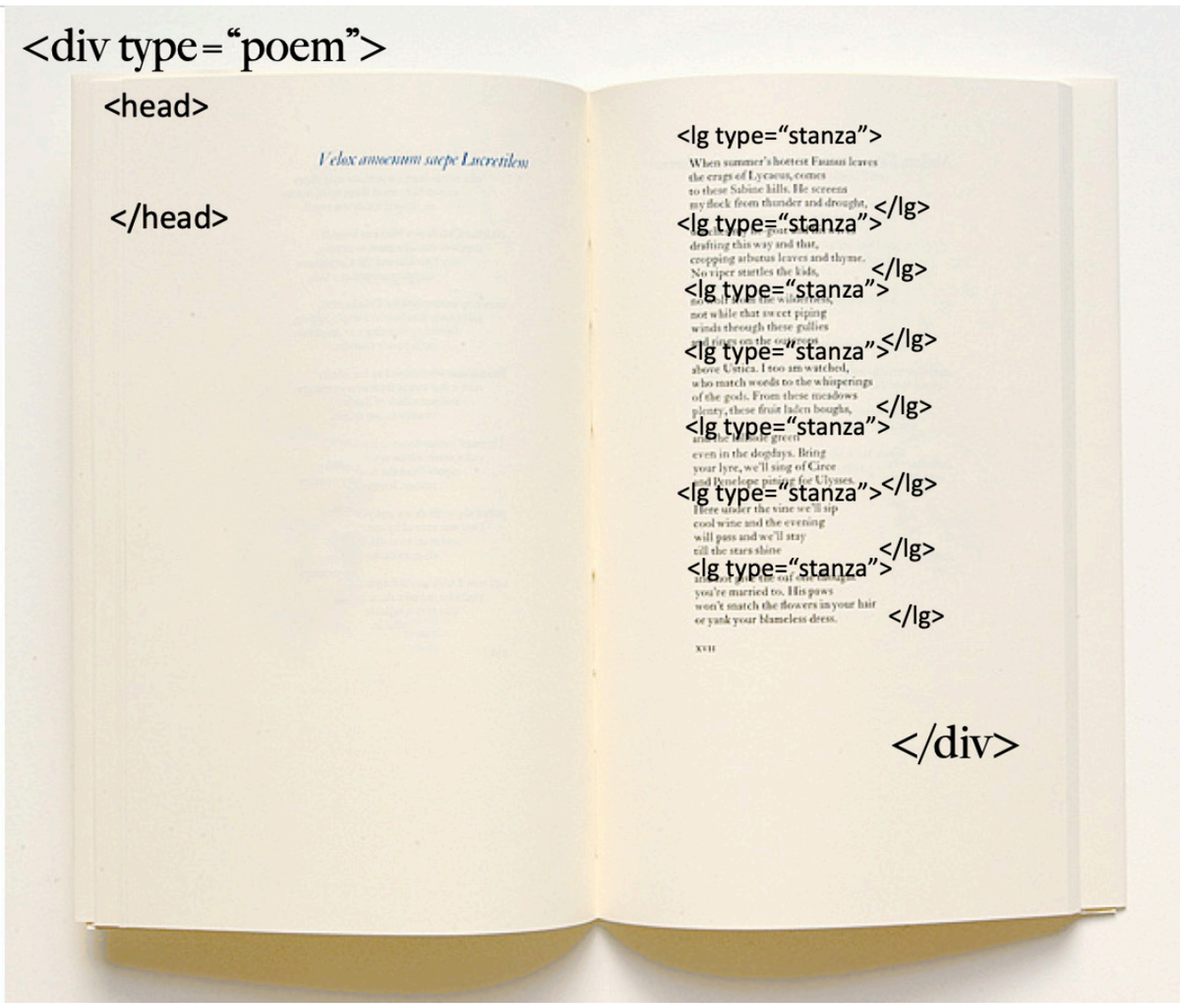


Figure 4: Branched mark-up of a poem

The above image[16] (Figure 4) is an example of a branched markup of a poem using necessary tags and elements like division type poem, line group stanza, line break and page break.

In the conceptual family tree of markup languages, SGML (Standard Generalised Markup Language) is often thought of as the parent of both HTML (Hypertext Markup Language) and XML (Extensible Markup Language). SGML is itself a descendant of IBM's GML (Generalized Markup Language). However, the support for the processing of SGML documents existed in relatively few applications, and specialism in learning SGML markup meant that it was not the universal answer for the markup of texts that many hoped it would become. XML, however, has become an international success in a very short span of time. It is now used across diverse academic and commercial domains—for documents, data files, configuration information, temporary and long-term storage, and for transmitting information locally or remotely, by both emerging start-ups and multinational conglomerates (Cummings 2013).

Consider an example of the earliest surviving Greek manuscript (Figure 5) of the old Bible from fourth century CE. The wonderful images on the British Library[17] show the present text in an unbroken flow of capital letters.

The case is not just with Greek; the picture depicted next is from an old manuscript from the fifteenth century by Al-Biruni, a scholar and polymath during the Islamic Golden Age[18]. The continuous pattern of writing Arabic (Figure 6) in three scripts (Naskh, Thalath and Kufi) in it.

Punctuation is a later addition to the language system to make it easier for non-native speakers or readers. Parkes in his *Pause and Effect: An Introduction to the History of Punctuation in the West* writes that '…copying

[16] Digital Editions: Start to Finish (an official course in the Programming4Humanists continuing education series) by Prof Laura Mandell (https://artsci.tamu.edu/english/contact/profiles/laura-mandell.html)

[17] https://blogs.bl.uk/digitisedmanuscripts/2016/12/explore-our-greek-manuscripts-online.html

[18] https://blogs.bl.uk/asian-and-african/2019/07/countdown-biruni-galileo-apollo-british-library-astronomy-manuscripts-in-new-visual-forms-.html

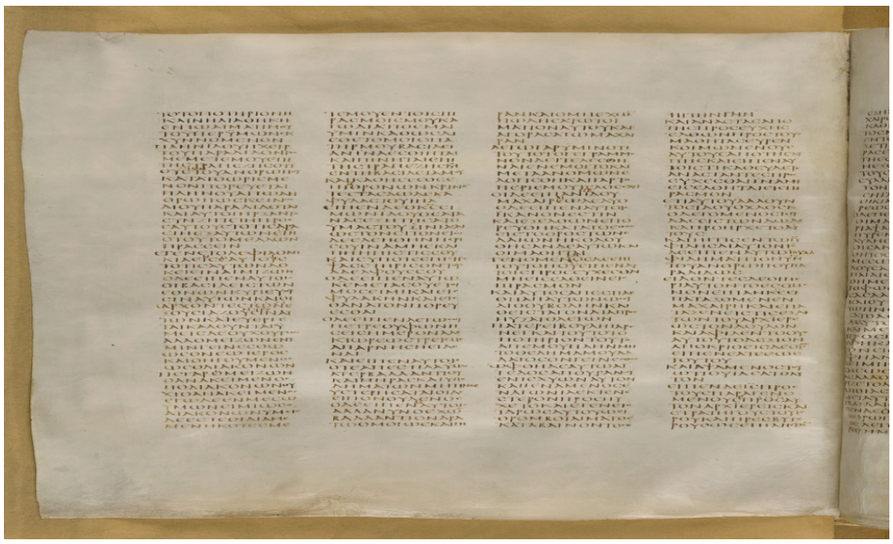


Figure 5: Greek manuscript

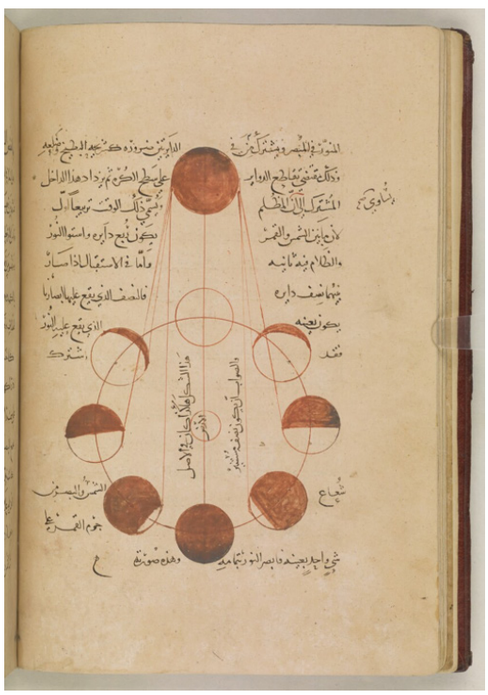


Figure 6: Arabic manuscript

of texts was an activity for inferior classes of people, who should not be encouraged to add to the text the level of interpretation that punctuation brings' (Parks 1992). However, these signs were not standardised. Therefore, we have different versions of the same text. For example, Quran in Arabic in Saudi Arabia is different from that in the subcontinent because of the added diacritics to ease the readability. Or for example, different versions of the Latin translation of the Bible, the Vulgate. And this led to the confusion in communication or while analysing the same text but with different versions. The increasing multiplication of texts created by printing came to drive an increased need for markup; this came from the need not only to refer analytically, as in a concordance, but also to refer between versions of texts. This, of course, implies the acceptance that a text can exist in versions, with variants; there is a concept of comparison here which may be profoundly subversive (Roueché 2012).

These three requirements, accessibility, longevity and integrity are the foundation of many of the intentions behind the creation of electronic scholarly editions and repositories of knowledge to disseminate them. Sadly, there are still many electronic editions produced to this day that fail to meet modern standards of accessibility; they often require a particular operating system and version of software (e.g., editions that function properly only in the Microsoft Internet Explorer web browser). This in turn jeopardises their longevity as, being

dependent on market forces for continued support, there is no guarantee that they shall continue to function in the future. This is slowly improving as funding bodies realise the need for proper accessibility and longevity from those resources produced with public money (Cummings 2013).

Selected Text: *Dastanbūy*

The project proceeds with acquiring the raw text of the chosen work, *Dastanbūy,* through an open-source web platform[19]. *Dastanbūy* is the personal diary of the renowned Indian Urdu poet, Mirza Asadullah Khan Ghalib. This book was initially composed in Persian and translated into English by Prof. Khwaja Ahmed Farooqi in 1970. The motivation for encoding an English translation of the Persian text in this digital version is that manually curating an XML file with right-to-left script is a laborious endeavour, particularly when incorporating tags. Engaging with right-to-left (RTL) scripts, such as Urdu, in XML-based platforms like oXygen introduces many methodological and technological obstacles for digital editors. Many XML editors are tailored for left-to-right languages, resulting in a visually perplexing arrangement of mixed-direction text, where Urdu content and XML tags coexist. Tag placement frequently becomes erratic, and the bidirectional characteristics of Urdu, which incorporate embedded Latin characters and numerals, may lead to improper text rendering or displacement of punctuation marks. Furthermore, search and validation operations are hindered by imperceptible directionality cues that disrupt XPath queries or schema validation. The restricted font and ligature support for Urdu significantly impedes reading during encoding, especially in poetic compositions.

Dastanbūy is a personal account of the Indian revolt of 1857 as witnessed by Ghalib and how it affected him severely as a court poet of the then ruling emperor, Bahadur Shah Zafar. As a resident of the Delhi province, which was the centre of the attack, this account praises and sympathises with the British and does not take side with the Indian rebel. Speculations have been made that this account is deliberately partial in its approach because Ghalib was under close surveillance and needed the foreign rulers' support for his livelihood.

A digital edition of *Dastanbūy* as a TEI document will add to the secondary sources of the historical research of mutiny, a literary analysis of Persian at the brink of losing its royal importance or any such questions that need a communication among scholars of various areas (reiterating the three goals of publishing a digital surrogate as discussed in the beginning of this essay).

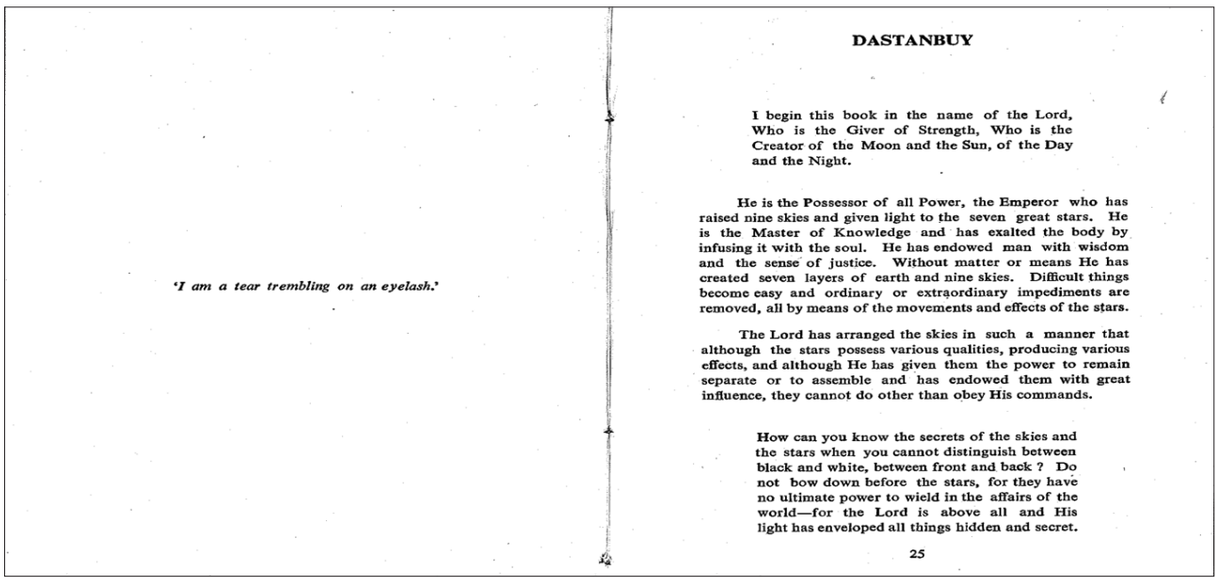
During the revolt, Ghalib wrote his diary of events called *Dastanbūy,* which means nosegay, in pure Persian with an unwitting admixture of Arabic words and in an oblique style of which he was a master and which the delicate occasion also demanded. The Persian text is translated by Prof. Khwaja Faruqi (Faruqi 1970) and covers the events of 15 months to the first of August 1858. This document serves as a diary, or more accurately, a chronicle of events primarily occurring in Delhi. However, when examined in chronological order, it represents a renewed effort to reaffirm his previous assertions regarding his pension and status, which he had relentlessly pursued since 1828 before the highest British officials in India and England. This is an attempt by Ghalib to absolve himself of involvement in the revolt of 1857, which ended in government by gallows, the blowing to bits of helpless multitudes, punishment-parades, the banishment of a whole population, and the hanging of many thousands of citizens after travesties of trial or none. Ghalib's *Dastanbūy* is important as it describes the story of the planned revolt, the ebb and flow of changing fortunes, of alternating hope and gloom as it affected a Delhi citizen, the throbbing of a sensitive soul and the reactions of a poet to an important historical situation, a story hitherto untold. This story has remained untold as it was impossible for Indians to tell it during those days of drumhead courts-martial, indiscriminate shootings and summary hangings. In the words of Vincent Smith, 'The story has been chronicled from one side only, and from one set of documents; or from no documents at Ghalib was writing under tremendous limitations. A slight suspicion would have cost him his life. Therefore, he has suggested the story rather than described it and has enhanced the effect of concealment by employing an oblique and formalised style and using obsolete words of pure Persian (Faruqi 1970).'

The material was readily accessible online through the digital collection of significant Urdu writings curated by Francis Pritchett[20], available as a PDF (Figure 7) of scanned photographs of the book, which was utilised as a key source for the research. The scanned images of the text were insufficient for the production of a TEI edition; raw text was necessary for encoding into XML structure and embedding with tags to enhance the text's structural and semantic richness. Consequently, to conserve time, energy and financial resources that would be unwisely expended on transcribing the full document (about 14,000 words), we attempted to utilise many accessible digital tools for optical character recognition, such as Tesseract or Google OCR.

The image above is a screenshot from the original source, a PDF file containing scanned photos of the document, which was to be divided into half (as it consists of a conjoined page) before employing the OCR tool to extract the text.

[19] https://www.ilovepdf.com/ocr-pdf

[20] https://franpritchett.com/00ghalib/texts/txt_dastanbu_kafaruqi.pdf



'I am a tear trembling on an eyelash.'

DASTANBUY

I begin this book in the name of the Lord, Who is the Giver of Strength, Who is the Creator of the Moon and the Sun, of the Day and the Night.

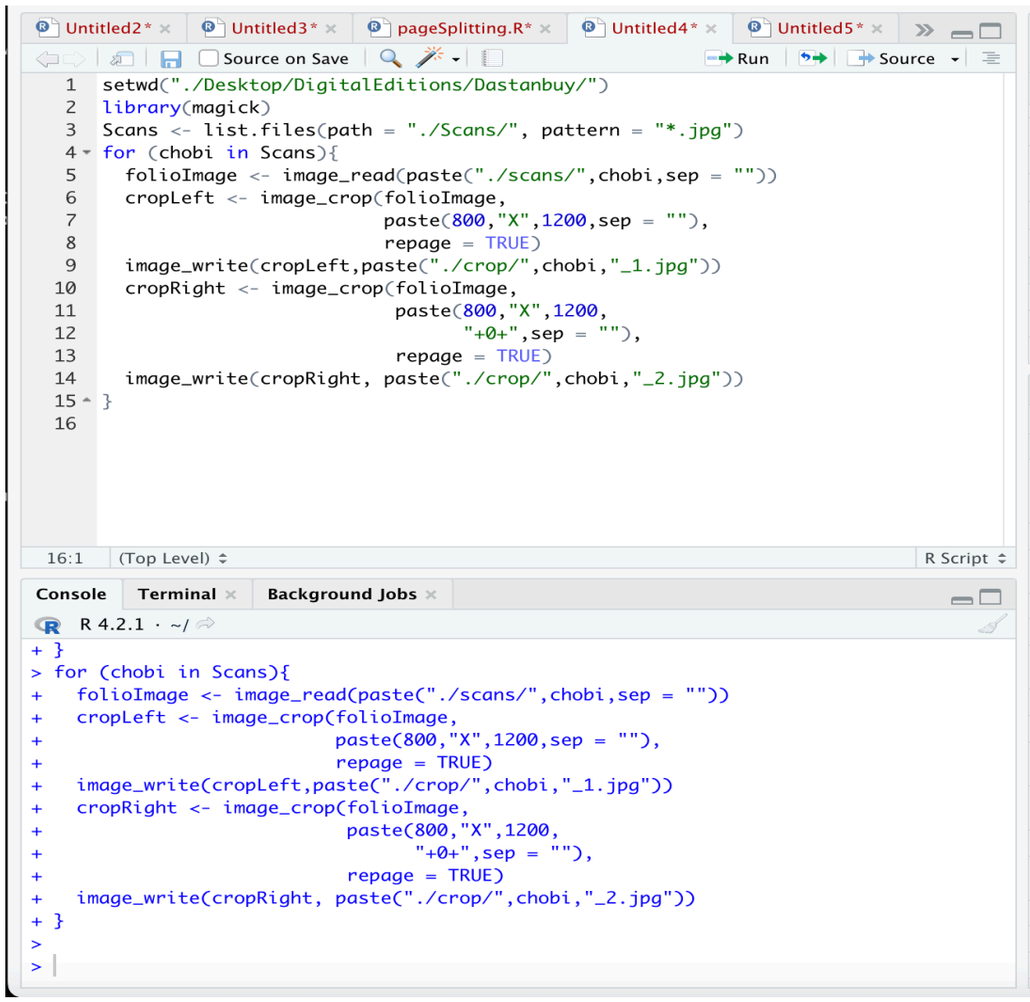
He is the Possessor of all Power, the Emperor who has raised nine skies and given light to the seven great stars. He is the Master of Knowledge and has exalted the body by infusing it with the soul. He has endowed man with wisdom and the sense of justice. Without matter or means He has created seven layers of earth and nine skies. Difficult things become easy and ordinary or extraordinary impediments are removed, all by means of the movements and effects of the stars.

The Lord has arranged the skies in such a manner that although the stars possess various qualities, producing various effects, and although He has given them the power to remain separate or to assemble and has endowed them with great influence, they cannot do other than obey His commands.

How can you know the secrets of the skies and the stars when you cannot distinguish between black and white, between front and back ? Do not bow down before the stars, for they have no ultimate power to wield in the affairs of the world—for the Lord is above all and His light has enveloped all things hidden and secret.

25

Figure 7: Scanned images PDF of Dastanbūy



Untitled2* | Untitled3* | pageSplitting.R* | Untitled4* | Untitled5*

Source on Save | Run | Source

```
1  setwd("./Desktop/DigitalEditions/Dastanbuy/")
2  library(magick)
3  Scans <- list.files(path = "./Scans/", pattern = "*.jpg")
4  for (chobi in Scans){
5    folioImage <- image_read(paste("./scans/",chobi,sep = ""))
6    cropLeft <- image_crop(folioImage,
7                           paste(800,"X",1200,sep = ""),
8                           repage = TRUE)
9    image_write(cropLeft,paste("./crop/",chobi,"_1.jpg"))
10   cropRight <- image_crop(folioImage,
11                           paste(800,"X",1200,
12                                 "+0+",sep = ""),
13                           repage = TRUE)
14   image_write(cropRight, paste("./crop/",chobi,"_2.jpg"))
15 }
16
```

16:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.2.1 · ~/

```
+ }
> for (chobi in Scans){
+   folioImage <- image_read(paste("./scans/",chobi,sep = ""))
+   cropLeft <- image_crop(folioImage,
+                          paste(800,"X",1200,sep = ""),
+                          repage = TRUE)
+   image_write(cropLeft,paste("./crop/",chobi,"_1.jpg"))
+   cropRight <- image_crop(folioImage,
+                           paste(800,"X",1200,
+                                 "+0+",sep = ""),
+                           repage = TRUE)
+   image_write(cropRight, paste("./crop/",chobi,"_2.jpg"))
+ }
>
>
```

Figure 8: Code in R for slicing scanned images

The preceding screenshot displays the code written in R (Figure 8) that was utilised for image cropping. Significant attention was devoted to tailoring the code to the document page's pixel dimensions. The following (Figure 9) summarises the instances of failure encountered while bisecting the page into two equal portions.

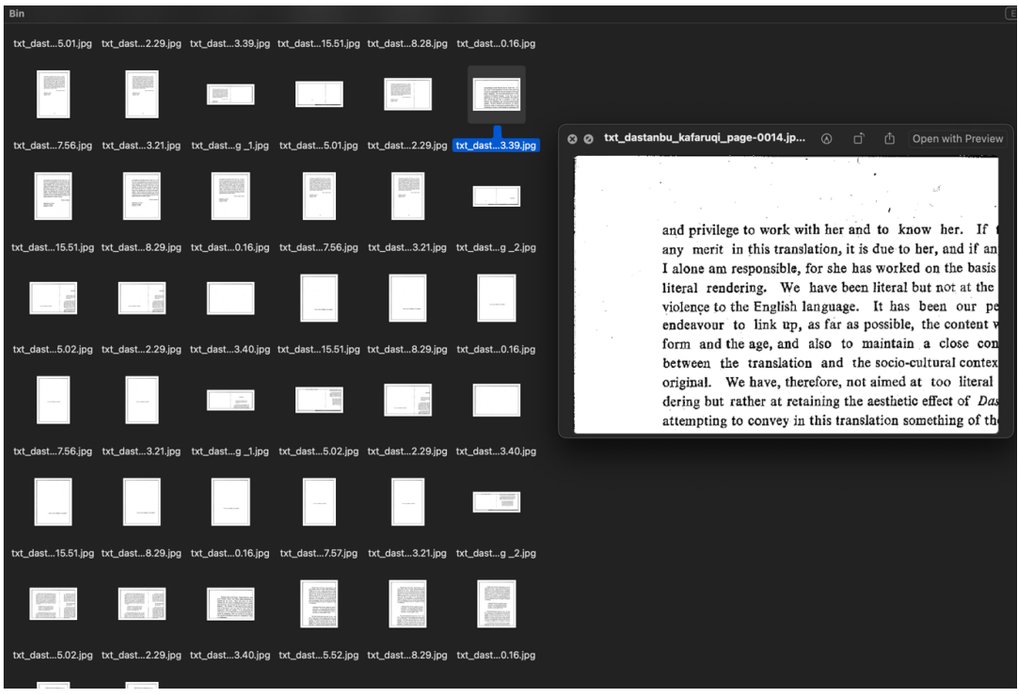


Figure 9: Instances of failure encountered

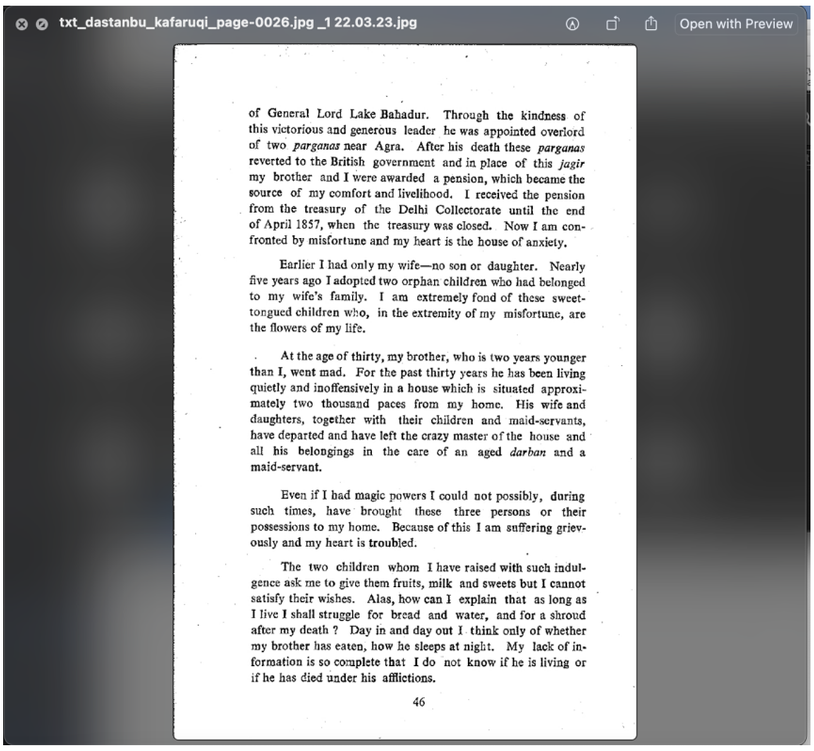


Figure 10: Final result of the successfully cropped page

Ultimately, upon the code functioning with accurate pixels, the document was successfully cropped using two methods:

1. Generating a single picture file that encapsulates the whole PDF content
2. Dividing every page into two halves

The following is the resulting cropped page (Figure 10). To avoid further delays on this ancillary activity, we utilised an online OCR engine[21] that yielded approximately 60 pages of complimentary OCR. The high quality of the scanned photos and the clarity of the

text enabled the OCR program to function effectively, resulting in a high standard of recognised text with little errors. OCR engines for English are highly advanced, resulting in superior text recognition quality compared to non-English languages.

The text extracted after using the online OCR tool was saved as a plain text file and was used to copy text and encode with necessary tags in the oXygen XML editor. It is an efficient and user-friendly XML editor and provides a comprehensive suite of XML authoring and development tools. It is available on multiple platforms, all major operating systems, and as a standalone application. oXygen XML Editor can be used in conjunction with all XML-based technologies, and it includes a large variety of powerful tools for creating, editing and publishing XML documents ('XML Editor', n.d.).

The process of encoding raw text with the available TEI tags is explained using the example of the poem *The Sick Rose* by William Blake, given below:

O Rose thou art sick.
The invisible worm,
That flies in the night
In the howling storm:
Has found out thy bed
Of crimson joy:
And his dark secret love
Does thy life destroy

XML is human-readable in the sense that its descriptive code consists of plain-text tags in angle brackets residing at the same level as the content they encode (that is, in the same document). These tags, which thus accompany the content wherever it goes, serve the same descriptive function as fields in a database (Liu, n.d.). In XML, for example, our poem might be marked up like so:

```
<div type="anthology">
    <poem><title>The SICK ROSE</title>
          <author>William Blake</author>
              <stanza>
                  <line>O Rose thou art sick.</line>
                  <line>The invisible worm,</line>
                  <line>That flies in the night</line>
                  <line>In the howling storm:</line>
              </stanza>
              <stanza>
                  <line>Has found out thy bed</line>
                  <line>Of crimson joy:</line>
                  <line>And his dark secret love</line>
                  <line>Does thy life destroy.</line>
              </stanza>
    </poem>
</anthology>
```

It is important to notice that the selected division type is an anthology, indicating that this segment of the markup poetry is part of a larger encoded anthology. There is a hierarchical arrangement of tags such that each tag must have a corresponding opening and closing at appropriate distinct levels; failure to adhere to this sequence throughout the file results in an error message. The title and author tags facilitate the identification of specific texts or entities such as countries, locations, first names and surnames. Stanza and line tags are utilised to preserve the poetic form in digital representations, closely resembling the original version.

The XML file of *Dastanbūy* comprises multiple tags essential for its presentation as a diary. The following is a list of the utilised tags in this paper as a minimal and introductory level of encoding:

1.
2. <hi rend="small caps">
3. <hi rend="large">
4. <hi rend="italics">
5. <p>
6. <lg type="stanza">
7. <lb>
8. <forename>
9. <nation designation>

Each tag serves a function that is visible in the final digital surrogate. In a TEI-encoded version, particular tags are employed to represent both the structure and visual attributes of a text. The tag indicates a specific portion of text and, with type="diary", designates a diary entry, enabling users to differentiate it from other text kinds. The <p> element marks paragraphs, whereas <lg type="stanza"> signifies a poetic stanza, maintaining the traditional structure of poetry. Line breaks within stanzas are denoted by <lb>. The <hi> tag emphasises text with certain visual attributes, such as rend="small caps", rend="large", or rend="italics", reflecting the author's or printer's intended emphasis. Names and entities are semantically annotated: <forename> denotes a personal first name, while <country> signifies nationality or ethnic classification. Collectively, these tags enable the edition to encapsulate both the textual content and its typographic, structural and semantic attributes in a manner that is machine-readable and interpretable for academic research.

Another important section is the TEI Header. One of the two mandatory sub-elements of the TEI document, the<tei header> includes all the metadata of your digital edition and information about the original source. It includes the names of original and digital surrogates, publication details, authorial details, web publishing server details, and the affiliation of the creator of digital edition. All this information must be encoded with particular tags and to be placed at proper positions. To simplify the process and save the manual work, an

online platform known as Replit[22] is available to create a TEI header with all your details. The user simply fills in all the details, and the platform creates a TEI header for your use in the XML file in the recommended format (Replit, n.d.). However, it is important to note that the header section is not visible in the digital edition available for public viewing. (Refer to line number 6 to 92 in the XML file to see the details included in the header section).

The prepared XML file was then converted into an XHTML format using an online tool called TEIGarage[23], provided by the TEI consortium on their official website. HTML and XHTML are two of the most popular markup languages used for developing web pages and applications. HTML is the standard markup language for creating web pages, while XHTML is a stricter and more standardised version of HTML. Both HTML and XHTML include a wide range of features, such as support for multimedia, styling and scripting. Here, we have preferred XHTML format, to which our final XML file has been converted.

stylesheets stored in CSS files. It is used to define styles for your web pages, including the design, layout and variations in display for different devices and screen sizes. W3Schools[24] is an online platform where different stylesheets are available and can be downloaded to add in your XHTML files ('CSS Introduction', n.d.). The third line of the XHTML file (refer to Figure 11) shows the linked stylesheets.

Finally, the XHTML file, along with its associated stylesheets, is hosted on GitHub webpage for public viewing. This TEI version serves as a publicly accessible sample intended for scholars unfamiliar with the notion of TEI. This is a concise digital scholarly edition or DSE that is semantically and structurally enriched by the incorporation of available tags. Readers are encouraged to experiment with TEI encoding in their projects by familiarising themselves with its principles and its applications for future endeavours. This paper's scope is confined to the creation of a publicly accessible TEI edition; nonetheless, these editions may be employed to analyse works pertaining to various research issues,

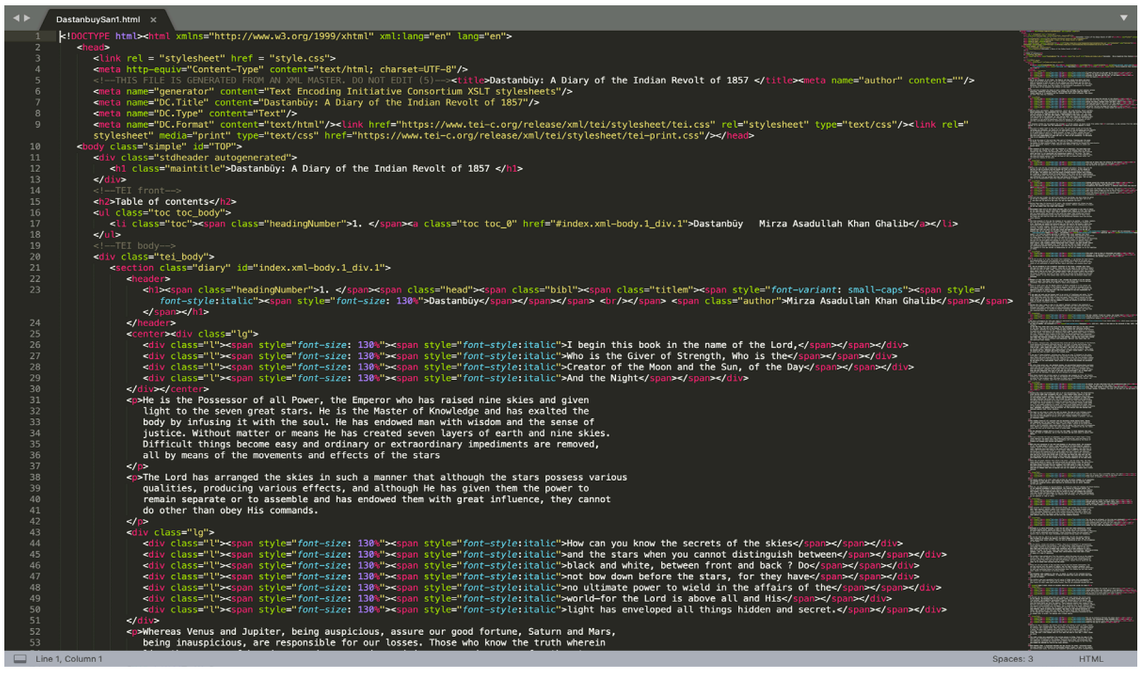


Figure 11: HTML with CSS

The above image (Figure 11) is a screenshot from Sublime Text, an editor that we used for adding the Cascading Style Sheets (CSS) files to the XHTML file for a beautiful rendition of our digital edition. CSS describes how HTML elements are to be displayed on screen, paper or in other media. It can control the layout of multiple web pages all at once, with external

which fall beyond the purview of this study and adhere to a distinct methodology.

The final digital edition with clickable links can be found here by following this link: https://johnnyblaze211.github.io/dastanbuy.html

The complete digital edition with other examples of literary texts are hosted on the following link: https://saniyairfan.github.io/SanDigEd.github.io/

[22] https://replit.com

[23] https://teigarage.tei-c.org

[24] https://www.w3schools.com

Discussion

Text Encoding Initiative is an advantage to the scholars of humanities and social sciences for sharing and analysing data together in a multi/interdisciplinary manner. The tags and annotations used in the digital surrogates can help navigate through a text and filter your search accordingly. If the study of literature is to become increasingly digital then we have an academic duty to ensure as much as possible that this is based on truly scholarly electronic editions that not only uphold the quality and reliability expected from such editions, but simultaneously capitalise upon the advantages that publication in a more flexible media affords (Cummings 2013).

Communication among different academic fields is necessary for all-round research. Markup languages and TEI facilitate this exchange of data in an efficient manner. When applied consistently, markup develops through the demands of people to communicate with one another, to discuss and understand texts. The determining essential is consistency—across kinds of material, languages and historical periods. This simple discipline is enabling us to communicate what we know and care about to new and unimagined audiences—just as the texts that survive on parchment or on stone are being read by generations their authors could never have imagined (Roueché 2012).

This study elucidates the methodology for constructing a fundamental SDE utilising TEI markup for beginners, accompanied by necessary resources. This paper does not target a heavily embedded file, rich in information or metadata; instead, it aims to familiarise readers with fundamental methods for utilising the TEI framework in their own customisable projects, employing numerous available tags and sources.

The development of a minimal DSE utilising the TEI framework underscores both a methodological breakthrough and an essential transformation in the conceptualisation, preservation and dissemination of South Asian humanities data. In contrast to descriptive digitisation methods that replicate texts as static images, TEI provides a semantic and interpretive framework that incorporates meaning, structure and scholarly intent inside the digital product. This method converts textual preservation into a process of knowledge modelling, delineating how the text is comprehended, classified and rendered interoperable across various disciplines and digital platforms.

The encoding of *Dastanbūy* demonstrates that TEI-based editions can harmonise cultural distinctiveness with universal norms. This study illustrates that utilising a minimum TEI schema, concentrated on structural and semantic components, can significantly improve discoverability, searchability and interpretive value. A modular and incremental model of TEI adoption enables local researchers to begin with limited resources, document their procedures, and expand collaboratively.

Moreover, the project emphasises that TEI is not simply a technical instrument but an essential technique. By prioritising openness, interoperability and citation, TEI positions the editor's interpretive work inside a reproducible framework, thereby connecting classical philology with computational analysis. A TEI edition serves as both an academic argument and a dataset, facilitating reuse by linguists, historians and literary researchers. Researchers aiming to engage in analogous endeavours should note that a streamlined TEI workflow provides an approachable gateway into digital humanities while maintaining academic rigour. Commencing with fundamental structural encoding, divisions, headers, entities and stylistic markup, scholars can create editions that are computationally advantageous, pedagogically enriching and sustainable for future inquiry. This methodology illustrates that significant digital editing can arise from constrained resources when driven by methodological rigour instead of technical excess.

This study establishes TEI as a conduit between preservation and interpretation, especially for endangered or geographically restricted literary traditions. As Indian digital humanities progresses, the use of TEI methodology will guarantee the preservation of texts and foster a culture of open, collaborative and critically informed digital study.

A forthcoming publication in the open access *Journal of the Text Encoding Initiative*[25] details the automation process for TEI markup and semantic annotation for extensive texts in languages other than English. Readers are advised to familiarise themselves with the framework presented in this paper and the referenced sources. They may advance to the subsequent level with the upcoming paper which talks in detail about the automated process of TEI and semantic markup using Large Language Models (LLMs) keeping human-in-the-loop approach. (Conference presentation slides are mentioned in the bibliography section on which this paper is based)

Acknowledgements

I wish to express my gratitude to Prof. Arjun Ghosh (Department of Humanities and Social Sciences, IIT Delhi) for his PhD-level course, 'Data in Humanities', and for recommending the course on TEI. This latter course, titled 'Fall 2022 Digital Editions: Start to Finish' is a continuing education course and online webinar, officially part of the Programming4Humanists series,

[25] Official journal of the TEI consortium which also publishes selected conference papers from the annual TEI event and can be accessed at: https://journals.openedition.org/jtei/

instructed by Prof. Laura Mandell, director of the Centre for Digital Humanities Research and professor in the Department of English at Texas A&M University. It was offered to me free of cost. Without their guidance, I would not have been able to discover the marvels of TEI and apply it in my work.

The data supporting this study are available in the repository linked as follows, with all the backend and frontend files: https://saniyairfan.github.io/SanDigEd.github.io/

Data availability statement. All data used in this study are publicly available at https://saniyairfan.github.io/SanDigEd.github.io/

Disclosure of Use of AI Tools. Generative AI tools were used to assist with code debugging and to provide minor suggestions related to grammar, spelling, and word choice. No AI tools were used beyond these limited purposes. The author has used only Quillbot for grammar check and proofreading.

Ethical standards. The research meets all applicable ethical guidelines. This study did not involve human participants, personal data, or animal subjects. The text used in this paper is publicly available.

Author contributions. The sole author was responsible for all aspects of this work.

Funding statement. A paid version of the software Oxygen XML Editor (https://www.oxygenxml.com) has been used in this project which was provided during a 4-month online workshop on Text Encoding Initiative, Fall 2022 Digital Editions: Start to Finish, an official course in the Programming4Humanists series organised by Centre of Digital Humanities Research and Department of Liberal Arts at the Texas A&M University, from September 30, 2022 to December 9, 2022.

Competing interests. The author declare no competing interests.

Reference List

Arko, Robert A., Kathryn M. Ginger, Kim A. Kastens, and John Weatherley. 2006. "Using Annotations to Add Value to a Digital Library for Education." *D-Lib Magazine* 12 (5). https://doi.org/10.1045/may2006-arko.

Bansode, Sadanand. 2008. "Creation of Digital Library of Manuscripts at Shivaji University, India." *Library Hi Tech News* 25 (1): 13–15. https://doi.org/10.1108/07419050810877508.

Cummings, James. 2013. "The Text Encoding Initiative and the Study of Literature." *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, John Wiley & Sons, Ltd, 2013, pp. 451–76. *DOI.org (Crossref)*, https://doi.org/10.1002/9781405177504.ch25.

CSS Introduction. https://www.w3schools.com/css/css_intro.asp. Accessed 19 November 2022.

Faruqi, Khwaja Ahmad, translator. 1970. *Dastanbūy: A Diary of the Indian Revolt of 1857 by Mirza Asadullah Khan Ghalib*. Asia Publishing House. https://franpritchett.com/00ghalib/texts/txt_dastanbu_kafaruqi.pdf

HTML vs XHTML: Know the Difference [2022 Edition] | Simplilearn. https://www.simplilearn.com/tutorials/html-tutorial/html-vs-xhtml. Accessed 19 November 2022.

Liu, Alan. 2004. *Transcendental Data: Toward a Cultural History and Aesthetics of the New Encoded Discourse*. p. 37.

Parks, M.B. 1992. *Pause and Effect an Introduction to the History of Punctuation in the West*. First. Routledge.

Sahle, Patrick. 2016. "What Is a Scholarly Digital Edition?" *Digital Scholarly Editing: Theories and Practices* 1:19–39.

Singh, Anil. 2012. "Digital Preservation of Cultural Heritage Resources and Manuscripts: An Indian Government Initiative." *IFLA Journal* 38 (4): 289–96. https://doi.org/10.1177/0340035212463139.

replit. 'Tei_header'. *Replit*, https://replit.com/@snowka/teiheader. Accessed 19 November 2022.

Roueché, Charlotte. 2012. 'Why Do We Mark Up Texts?' *Collaborative Research in the Digital Humanities*. Taylor and Francis Group.

Taj, Amreen, and Bhakti Gala. 2024. "Digitization Projects for Cultural Heritage Materials: A Study With Special Reference to Arabic, Persian, and Urdu Manuscripts." In Advances in Library and Information Science, edited by K. R. Senthilkumar. IGI Global. https://doi.org/10.4018/979-8-3693-2782-1.ch013.

TEI: Text Encoding Initiative. https://tei-c.org/. Accessed 19 November 2022.

TEIgarage: https://teigarage.tei-c.org. Accessed 11 June 2025.

XML Editor. https://www.oxygenxml.com/xml_editor.html.

Forthcoming paper's conference presentation slides: https://doi.org/0.5281/zenodo.13997483