Developing a TEI-based Approach for Encoding Premodern Islamic

Adrien de Jarmy, Clarck Junior Membourou Moimecheme

▶ **To cite this version:**

Adrien de Jarmy, Clarck Junior Membourou Moimecheme. Developing a TEI-based Approach for Encoding Premodern Islamic. 2025. ⟨hal-05302214⟩

HAL Id: hal-05302214

https://hal.science/hal-05302214v1

Preprint submitted on 7 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

1. **Article title: Developing a TEI-based Approach for Encoding Premodern Islamic Texts**

2. Authors: Adrien de Jarmy (Associate Professor, University of Strasbourg), Clarck Membourou Moimecheme (Postdoc, University of Strasbourg)

3. Abstract: This paper presents a TEI-based approach for encoding premodern Islamic texts, focusing on the development and application of a standardized TEI XML schema. Over the past decade, TEI XML models for Arabic have primarily addressed linguistic and grammatical issues, particularly in encoding Arabic lexicons and managing textual variations in manuscripts. However, the need for a comprehensive and consistent workflow for encoding premodern Islamic texts remains largely unmet. The Badr Project aims to fill this gap by proposing a TEI XML text database that covers sources up to the 14th century. This database emphasizes key Islamic texts structured around the hadith format, linking the content of a text (*matn*) to a chain of guarantors (*isnād*). The paper outlines the rationale behind the selection of tags and attributes used in encoding, describes the semi-automatic encoding process, and explores the research possibilities enabled by this method. The proposed approach facilitates the extraction and quantification of named entities, thereby opening new avenues for the study of premodern Islamic texts and historiography.

4. Keywords: Islamic studies; Premodern Islamic Texts; Hadith; Digital humanities; TEI XML; Named Entity Recognition.

5. Text of the article: see next page and p. 18 for biographical statements and project informations

Introduction

For over a decade, TEI XML models for Arabic have focused on several key areas, including linguistic and grammatical issues, particularly the encoding of Arabic lexicons. This encompasses research on morphological encoding, the preservation of syntactic variety (Nahli and Del Grosso 2021 and 2020; Maraoui and Haddar 2015), and lexicon creation for both Standard and Classical Arabic (Olivieri 2018; Olivieri, Pepe and Cicola 2016). Additionally, driven by the movement of digitization of manuscripts, TEI standards have been applied to the editing of Arabic manuscripts and the management of textual variations (Brinis, Soualah and Bouarab 2023; Clivaz and Schulthess 2017; Lancioni and Joose 2016; Salah and Hassoun 2012). However, TEI is not the only approach to encoding Arabic texts from this perspective. Members of the KITAB project, which aimed to create a large, open-access, machine-readable collection of premodern Arabic and Persian texts — known as the *Open Islamicate Text Initiative* (OpenITI) — developed a lightweight tagging system based on Markdown. This adaptation was designed specifically to facilitate the large-scale study of text reuse across extensive corpora[1]. The KITAB team employs Passim, an advanced algorithm for detecting textual reuse through the comparison of 300-word segments between pairs of documents in Arabic[2]. Similarly, the team of the *Kalīla and Dimna: AnonymClassic* project at Freie Universität Berlin, which seeks to create a digital synoptic edition of the *Kalīla wa-Dimna* fables[3] by tracing their transmission across nearly one hundred Arabic manuscripts, also chose not to use TEI XML. They cited its inefficiency with large corpora, steep learning curve, and lack of specificity for encoding Arabic manuscript tradition (Kozae 2022).

If these two projects decided to move away from TEI, what benefits can TEI still offer in the field of Arabic studies and beyond? Another area of study focuses on the encoding of OCR-processed texts, not for editorial purposes, but for data extraction — such as Named Entity Recognition (NER) — to facilitate indexing and historical analysis. In this context, TEI XML is particularly useful for moderately sized corpora of texts, typically ranging to several tens of thousands of words. The aim here is not to process extreme vast amounts of data but to statistically evaluate the distribution of named entities within the texts. These entities can include various characters, authorities, locations, objects, and other significant references. While some efforts have been made to begin encoding ancient Islamic texts, existing work has typically focused on specific sources or genres (Cruse and Sajawel 2024; Maroui, Haddar and

[1] https://kitab-project.org/about/.
[2] https://kitab-project.org/methods/text-reuse.
[3] https://www.geschkult.fu-berlin.de/en/e/kalila-wa-dimna/index.html.

Romary 2018 and 2017). This highlights the absence of a standardized TEI XML schema for encoding and a consistent workflow for its implementation.

One of the key objectives of The Badr Project, a research initiative funded by the French Institute for Islamology (IFI)[4], is to propose a solution to fill this gap. The project was initially built around the idea of studying the construction, development, and transmission of various narratives surrounding the Battle of Badr. This pivotal early Islamic battle, fought in 2 AH (624 CE), saw the Prophet Muḥammad and his Companions, based in Medina, confront the Meccans. The project includes the creation of a comprehensive TEI XML text database covering sources up to the 14th century. This database highlights key Islamic texts using the hadith structure, where the content of a text (*matn*) is linked to a chain of guarantors (*isnād*). From the 14th century onward, the project will also examine the evolving interpretations of the battle, culminating in a symposium that will bring together specialists in Strasbourg in November 2025[5]. The corpus comprises 40 texts from various "genres", including historiography (*Sīra*, *maghāzī*), biographical dictionaries (*ṭabaqāṭ*), hadith collections with a jurisprudential perspective (*fiqh*), and Qurʾānic commentaries (*tafsīr*), in which references to the battle vary in prominence. The texts range in length from a few thousand to tens of thousands of words. Reference to the battle of Badr can take the shape of full chapters dedicated to the event, especially in historiography, or are scattered references across the text.

The first part of this paper will present the rationale behind the selection of tags and attributes used to encode premodern Islamic texts structured around hadith transmission. The second part will outline the workflow, including the semi-automatic encoding process and the extraction and quantification of named entities. Finally, the paper will explore the kinds of research questions this method opens up for the study of premodern Islamic texts and historiography.

[4] https://islamologie.unistra.fr/actualites/seminaire-de-recherche-islamologie-et-humanites-numeriques/.
[5] https://islamologie.unistra.fr/actualites/appel-a-contributions-pour-le-colloque-international-badr-ecriture-et-memoire-de-la-bataille-de-badr-viie-xxie-siecle/.

Encoding choices and rationale

teiHeader and metadata guidelines

To illustrate the structure of the teiHeader, we will use as an example an edition of the *Sīrat al-nabawiyya* by Ibn Hishām (d. 833), one of the most renowned biographical accounts of Muḥammad, the Prophet of Islam, compiled during the early Abbasid period.

The **<fileDesc>** contains all the essential bibliographic information. It is subdivided into several key elements.

The tag **<titleStmt>** specifies the original title of the work and identifies both the author and the modern editor. In our example:

```
<titleStmt>
<title/>السيرة النبوية لابن هشام<title>
<author>
<persName/>عبد الملك بن هشام بن أيوب الحميري المعافري أبو محمد جمال الدين<"author"=type persName>
<author/>
<editor>
<persName/>مصطفى السقا<"editor"=type persName>
<editor/>
</titleStmt>
```

Fig 1. <titleStmt>

The inclusion of **<persName>** elements for both author and editor, along with the @type attribute, allows to distinguish historical authorship from modern editorial intervention. This distinction is crucial when working with classical Arabic texts, whose printed editions often reflect complex histories of compilation, redaction, and transmission.

The tag **<publicationStmt>** records the publication details of the printed edition used as the base for encoding:

```
<publicationStmt>
<publisher/>شركة مكتبة ومطبعة مصطفى البابي الحلبي وأولاده بمصر<publisher>
<pubPlace/>Cairo<pubPlace>
<date/>1955<date>
</publicationStmt>
```

Fig 2. <publicationStmt>

And **<editionStmt>** indicates the edition number:

```
<editionStmt>
<edition/>2<edition>
</editionStmt>
```

Fig 3. <editionStmt>

Withing **<sourceDesc>**, ** ** is used to describe the scope of encoding within a source. In this case, we specify that the XML file encodes only a portion of the full work—specifically the chapter on the Battle of Badr in the text of Ibn Hishām:

This encoding represents only the chapter on the Battle of Badr (غزوة بدر الكبرى)

A particularly important modeling decision concerns the use of nested ** ** elements with @type attributes to capture the layered and heterogeneous structure of the text. We can distinguish among:

- chapter
- subchapter
- isnad (chains of transmission)
- matn (main narrative body)
- fiqh (jurisprudence, legal discussions)
- poetry (or *shaʿr* in Arabic)
- quran (with @n for sura and verse numbers)
- tafsir (exegetical texts, also known as *tafsīr* literature)
- bible (biblical references)

This typology is intended to reflect the textual layering characteristic of premodern Islamic texts, particularly hadith literature. ** ** elements may be used within the **<sourceDesc>** to indicate the overall genre of the encoded text. Alternatively, they can appear within the body of the text to mark specific sections that belong to a particular genre or textual stratum. **<head>** is also used to encode the title of a chapter or subchapter. For instance:

```
                                        <"chapter"=type div>
                                    <head/>غزوة بدر الكبرى<head>
                <div/>[any list of authorities]<"isnad"=type div>
<div/>[text associated with the former isnād]<"matn"=type div>
        <div/>[legal discussion, jurisprudence]<"fiqh"=type div>
                                                        <div/>
```

Fig 4. Use of elements

The metadata structure also anticipates the presence of structured lists**,** common in premodern Islamic texts, such as those enumerating participants, animals, objects, and prisoners. These are encoded with **<list>** elements, again qualified by a @type attribute. Likewise, dates are normalized using the **<date>** tag with @when attributes.

Named entity encoding guidelines

The encoding of named entities follows TEI standards, adapted to the specificities of premodern Islamic texts. The aim is to enable precise semantic indexing and facilitate structured querying and analysis of individuals, places, objects, and conceptual terms within the corpus.

<persName> is used to tag all proper names referring to human individuals:

- **@type** indicates the role or classification of the person:
    - author: individuals credited with authoring a text, ex: Ibn Hishām.
    - editor: individuals responsible for editing or transmitting a version of a text.
    - authority: figures cited in *isnād*-s or chains of transmission, ex: Ibn Isḥāq (m. 150/767), the main authority quoted in Ibn Hishām's *Sīra*.
    - prophet: reserved for Muḥammad; also applies to expressions such as *rasūl Allāh* (رسول الله) or *nabī* (نبي).
    - sahabi: Companions of Muḥammad.
    - tabiun: members of the generation following the Companions.
    - Additional subtypes may be specified, such as:
        - mekkan: figures opposed Muḥammad from Mecca.
        - muhajir[6], ansar[7], caliph, mawla[8], etc., with cumulative values (ex: type="sahabi muhajir caliph").
- **@state** describes the status of the person in the text (ex: dead, martyred…).
- **@xml:id** is used to assign a unique identifier to each individual, allowing for consistent referencing across the corpus.

The tag **<orgName>** is used to encode names of collective entities such as tribes, religious communities, or political groups:

- **@type**:
    - tribe: for Arab tribal affiliations, ex: Quraysh (قريش), the tribe of Muḥammad.
    - religion: for religious communities, ex: Jews (يهود).
- **@xml:id** again ensures unambiguous identification of named groups across documents.

<placeName> is used for all geographical and architectural entities:

[6] Most commonly, a Compagnion of Muḥammad who emigrated from Mecca to Medina, although the definition of the concept evolved with time.
[7] A Companion who welcomed Muḥammad and his followers from Mecca in the city of Medina.
[8] Freed slave.

- **@type** classifies the place: city, landscape, building, etc.
- **@subtype** refines this classification:
    - For natural features, ex: mountain (جبل), valley (وادي), river (نهر), etc.
    - For buildings, ex: mosque (مسجد), well (بئر), etc.
- **@xml:id** enables stable referencing of proper names, ex: xml:id="well_of_Badr" for بئر بدر.

<object> tags named or unnamed objects:

- **@type** indicates the general category, ex: weapon, clothe.
- **@subtype** specifies the object:
    - For weapons, ex: sword (سيف), shield (درع), spear (ربة), bow (قوس), etc.
    - For garments, ex: turban (عمامة), etc.
- **@xml:id** assigns identifiers to specific named objects, ex: xml:id="dhu_l-fiqar" for the Muḥammad's sword.

Non-human entities and descriptive terms that do not fall under persons, places, or objects are encoded using the tag **<rs>** (referencing string):

- **@type** indicates the semantic field:
    - animal, ex: camel (جمل), horse (حصان), snake (ثعبان), fish (سمكة), etc.
    - color, shape, for physical descriptions.
    - surnatural, for supernatural beings, ex: angel (ملاك), djinn (جن), devil (شيطان), etc.
    - Virtue, emotion: for moral or affective qualities, ex: strength (قوة), worry (قلق), etc.
    - Other types include body, transport, fiscal, war, hereafter, food, manner, ritual, rhetoric, plants, family.
- **@subtype** is used where further specificity is required.
- **@xml:id** is reserved for named non-human entities, ex: the angel Gabriel or *Jibrīl* (جبريل).

Semi-Automated Encoding and Named Entity Extraction

Concomitantly with the development of the encoding guide, questions also arose regarding technological choices related to software tools and the methodology to be adopted in the procedures for the standardisation of texts, tagging, annotation, normalisation, extraction,

and structuring of data. The decision to prioritise the use of open-access tools and resources allowed for consistency with the open scientific approach underpinning the research conducted within the Badr project. The selected tools, freely accessible online, were chosen for their operability and advanced functionalities, which are compatible with the TEI-XML markup language, thereby contributing to the efficient and structured processing of data derived from the Arabic-language corpus.

The Manual Encoding Work on Notepad++

Following the acquisition of Arabic-language texts in .txt format, the first step consisted of manually encoding the historical Arabic texts. This task was carried out using the free source code editor Notepad++[9], with strict reference to the guide outlined in the preceding pages:

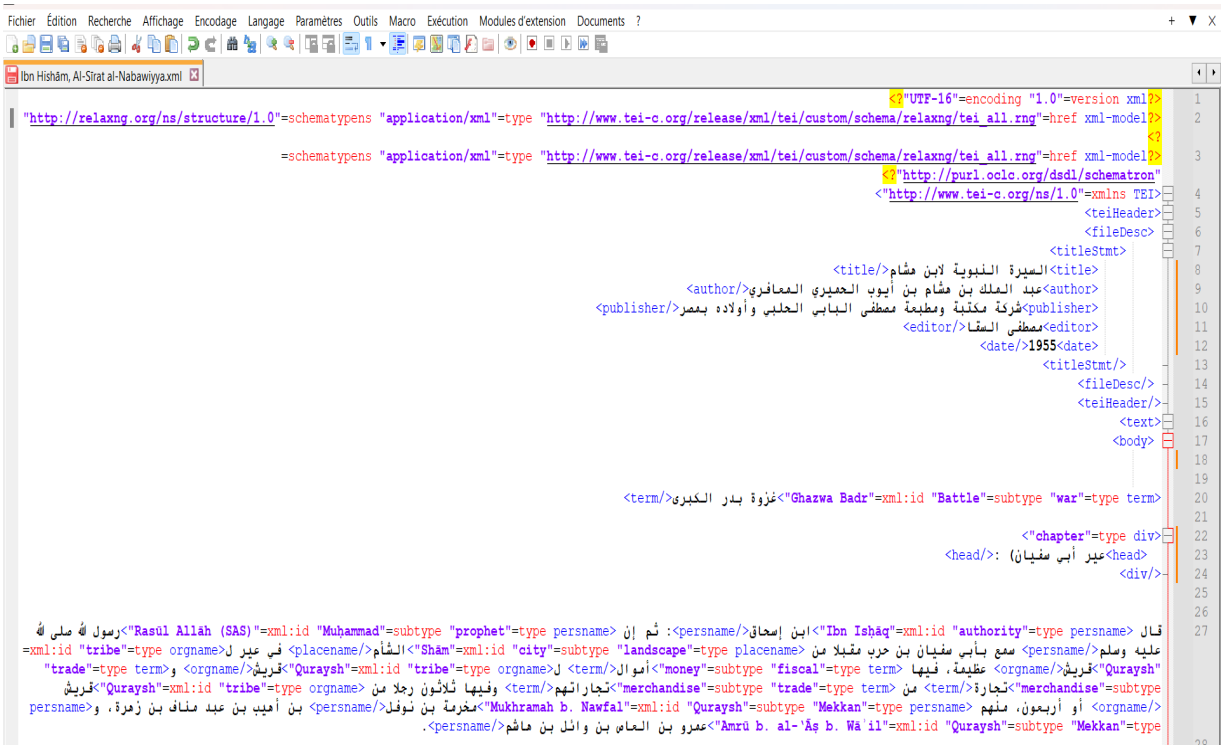


Fig 5. Beginning of the encoding of Ibn Hishām's *Sīrat al-nabawiyya* in the Notepad++ interface

This open-source software, developed by the computer scientist Don Ho, is available for the Windows environment. Written in C++, it features a classic graphical interface, high execution speed, and functionalities (such as syntax highlighting) that enhance productivity. Arabic is supported by Notepad++, which enabled effective management and processing of right-to-left (RTL) script during the tagging of entities in the texts.

[9] https://notepad-plus-plus.org/

Development Workflow for the Script in Visual Studio Code

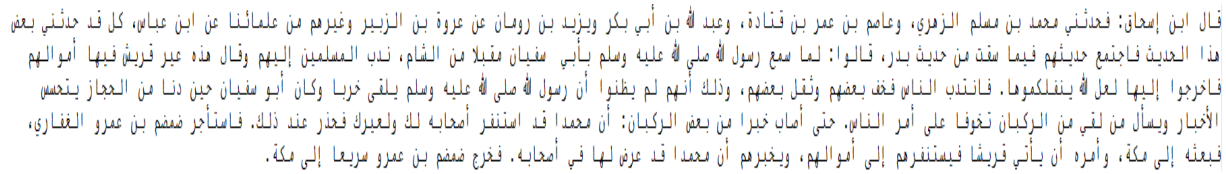
This stage allowed the manual tagging preparation work carried out in Notepad++ to be put to use. The creation of a dataset proved valuable for the process of automating the encoding of texts. Indeed, this approach follows an ETL (extract, transform, load) logic, a robust method widely used in data processing and loading (Kimball, 2004). The automatic encoding task required the implementation of three scripts[10]: the manual tag learning script, the automatic tagging script, and the data visualisation script.

Script n°1: Manual Tag Learning

Based on the dataset produced from manual encoding, this learning script processes XML files already encoded manually in accordance with the TEI standard. It is therefore essential for training models capable of correctly identifying and categorising named entities in Arabic-language texts. Using Python's XML module for parsing, named entities tagged according to the TEI-XML standard are traversed and stored in a database. Each newly encoded entity is then compared to the existing records in the database, and if it is absent, it is automatically inserted. This method enables supervised incremental learning, allowing the database to be progressively enriched as new documents are processed.

Script n°2: Automatic Tagging

This script enables the automatic application of tags to named entities in new texts. This process relies on the entities and attributes already recorded during the learning phase. The algorithm identifies known entities (persons, organisations, places, etc.) and surrounds them with XML tags. The generated tags are inserted into a complete TEI structure, including headers, schema declarations, and metadata. The tagged documents are thus structured and immediately usable for text mining, in compliance with technical standards, while ensuring the interoperability of the corpus.


قال ابن إسحاق: فحدثني محمد بن مسلم الزهري، وعاصم بن عمر بن قتادة، وعبد الله بن أبي بكر ويزيد بن رومان عن عروة بن الزبير وغيرهم من علمائنا عن ابن عباس، كل قد حدثني بعض هذا الحديث فاجتمع حديثهم فيما سقت من حديث بدر، قالوا: لما سمع رسول الله صلى الله عليه وسلم بأبي سفيان مقبلا من الشام، ندب المسلمين إليهم وقال هذه عير قريش فيها أموالهم فاخرجوا إليها لعل الله ينفلكموها. فانتدب الناس فخف بعضهم وثقل بعضهم، وذلك أنهم لم يظنوا أن رسول الله صلى الله عليه وسلم يلقى حربا وكان أبو سفيان حين دنا من الحجاز يتحسس الأخبار ويسأل من لقي من الركبان تخوفا على أمر الناس، حتى أصاب خبرا من بعض الركبان: أن محمدا قد استنفر أصحابه لك ولعيرك فحذر عند ذلك، فاستأجر ضمضم بن عمرو الغفاري، فبعثه إلى مكة، وأمره أن يأتي قريشا فيستنفرهم إلى أموالهم، ويخبرهم أن محمدا قد عرض لها في أصحابه. فخرج ضمضم بن عمرو سريعا إلى مكة.

[10] To consult the scripts, see https://gitlab.huma-num.fr/mantonio/badr.

<persname/>محمد بن مسلم<"Muḥammad b. Muslim"=xml:id "authority"=type persname> فحدثني :<persname/>ابن إسحاق<"Ibn Isḥāq"=xml:id "authority"=type persname> قال
الزهري، و<persname/>عاصم بن عمر بن قتادة<"Āṣim b. 'Umar b. Qatāda"=xml:id "authority"=type persname>،
<"Abd Allāh b. Abī Bakr"=xml:id "authority"=type persname>عبد الله بن أبي بكر<persname/> و
<"Urwa b. al-Zubayr"=xml:id "authority"=type persname> عن <persname/>يزيد بن رومان<"Yazīd b. Rūmān"=xml:id "authority"=type persname>
عروة بن الزبير<persname/> وغيرهم من <persname/>علمائنا<"Ulama"=xml:id "authority"=type persname> عن
<"Ibn 'Abbās"=xml:id "authority"=type persname>ابن عباس
<"Rasūl Allāh (SAS)"=xml:id "Muḥammad"=subtype "prophet"=type persname>، كل قد حدثني بعض هذا الحديث فاجتمع حديثهم فيما سقت من حديث بدر، قالوا: لما سمع
رسول الله صلى الله عليه وسلم<persname/> بأبي سفيان مقبلا من <placename/>الشام<"shām"=xml:id "city"=subtype "landscape"=type placename>، ندب ال
<orgname/>مسلمين<"muslimīn"=xml:id "religion"=type orgname> إليهم وقال هذه عير <orgname/>قريش<"Quraysh"=xml:id "tribe"=type orgname> فيها
<term/>أموال<"money"=subtype "fiscal"=type term>
هم فاخرجوا إليها لعل الله ي<term/>نفل<"war booty"=subtype "fiscal"=type term> كموها. فانتدب الناس فخف بعضهم وثقل بعضهم، وذلك أنهم لم يظنوا أن
<persname/>رسول الله صلى الله عليه وسلم<"Rasūl Allāh (SAS)"=xml:id "Muḥammad"=subtype "prophet"=type persname> يلقى حربا وكان أبو سفيان حين دنا من
<placename/>الحجاز<"Ḥijāz"=xml:id "region"=subtype "landscape"=type placename> يتحسس الأخبار ويسأل من لقي من الركبان تخوفا على أمر الناس. حتى أصاب خبرا من بعض الركبان: أن محمدا قد استنفر أصحابه لك ولعيرك فحذر
عند ذلك. فاستأجر <persname/>ضمضم بن عمرو الغفاري<"Ḍamḍam b. 'Amrū"=xml:id "Mekkan"=type persname>، فبعثه إلى <orgname/>قريش<"Quraysh"=xml:id "tribe"=type orgname>! فيستنفرهم إلى <term/>أموال<"money"=subtype "fiscal"=type term>هم،
وأمره أن يأتي <placename/>مكة<"Makka"=xml:id "city"=subtype "landscape"=type placename> سريعا إلى <persname/>ضمضم بن عمرو<"Ḍamḍam b. 'Amrū"=xml:id "Mekkan"=type persname> فخرج
ويخبرهم أن محمدا قد عرض لها في أصحابه.
.<placename/>مكة<"Makka"=xml:id "city">

Fig 6. Automatically encoded raw text excerpt from the *Sīrat al-Nabawiyya* by Ibn Hishām

Script n°3: Data Visualization

The third script enables the creation of visualisations using powerful libraries to display and analyse data intuitively. The Matplotlib library is employed to generate a variety of charts, thereby facilitating the presentation of information in a visual format[11]. Meanwhile, Pandas plays a key role in data management by organising data into DataFrames, a structure that allows for efficient manipulation and representation of data prior to visualisation[12]. By combining these two tools, the script offers a comprehensive solution to transform raw data into clear and informative graphics.

This third script facilitates the creation of visualisations for macro-level analyses, thereby enabling a comprehensive statistical overview of the encoded named entities subsequently stored in the database:

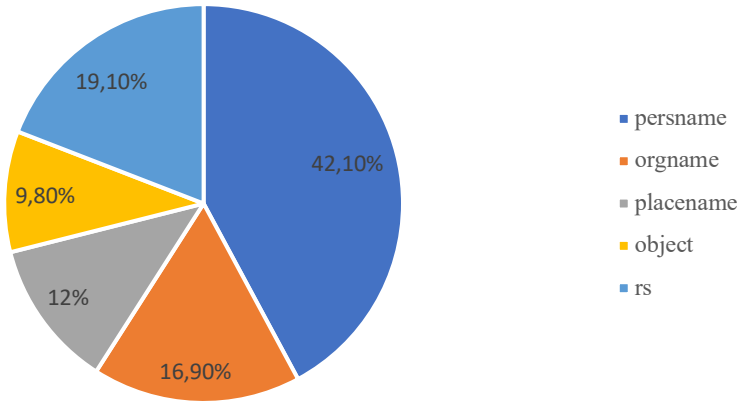


Fig 7. Percentage of encoded named entities

[11] https://matplotlib.org/.
[12] https://pandas.pydata.org/.

This pie chart visualisation provides a key to macro-level analysis of the data stored in the database. The extraction of named entities first allows for a prosopographical analysis, identifying major or marginal figures involved in the Battle of Badr, whether belonging to the camp of the allies or the Prophet's enemies. The strong centrality of the human actors in the conflict facilitates the construction of a co-occurrence network between individuals mentioned in the accounts (tag *persname*, 42%) and their collective designations (tag *orgname*, 16.9%). Thus, examining the number of occurrences and the distribution of the qualifiers used for the two opposing groups at the Battle of Badr allows for the formulation of several hypotheses, which will need to be confirmed or refuted upon completion of the full corpus encoding.

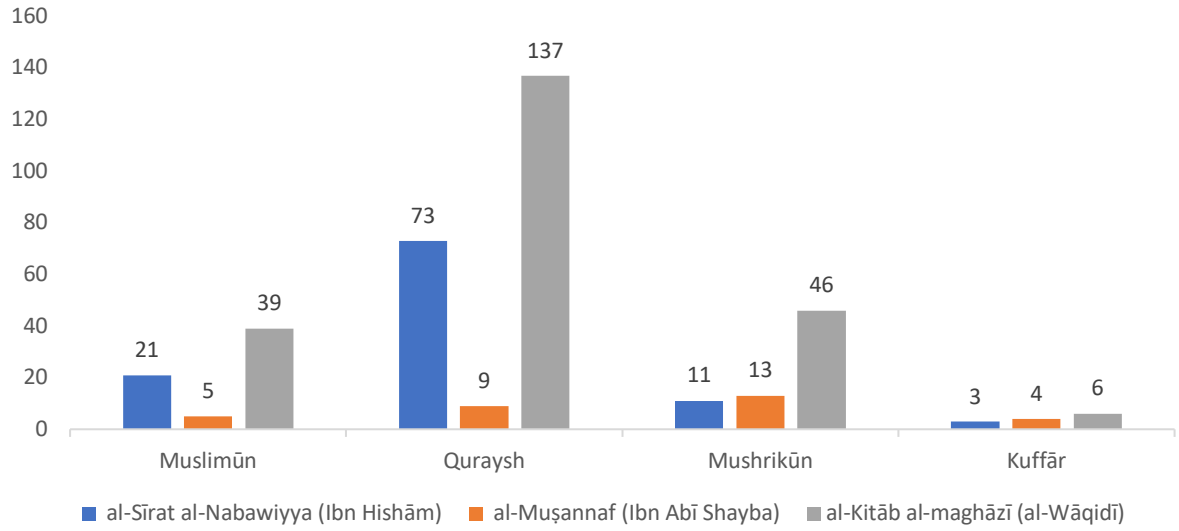


Fig 8. Distribution of the Descriptive Epithets Attributed to the Two Belligerent Camps during the Battle of Badr

Queries conducted across three texts (Ibn Abī Shayba, *al-Muṣannaf*; Ibn Hishām, *Sīrat al-Nabawiyya* and al-Wāqidī, *Kitāb al-Maghāzī*) reveal, on the one hand, the use of a quasi-Qur'anic terminology that contrasts the associators (*mushrikūn*) or non-believers (*kuffār*) with the believers (*muslimūn*). It is evident that this normative terminology predominates in particular in Ibn Abī Shayba's *al-Muṣannaf*, where the vocabulary is consistent and serves to clearly define who was on the "right" or "wrong" side during the battle.

On the other hand, there is a marked increase in the use of the term Quraysh, notably in Ibn Hishām's *Sīrat al-Nabawiyya* and al-Wāqidī's *Kitāb al-Maghāzī*, indicating that the enemies of Muhammad are largely identified as members of this Meccan tribe. In the *Maghāzī* texts, this dominance of tribal terminology suggests that Badr is primarily perceived through the lens of genealogy. This is partly due to the nature of these texts, which, being independent from the

hadith collections, include genealogical lists of the combatants. These lists serve as reference points to establish whether an individual's ancestor took part in the battle, on which side, and at what moment, thereby giving the memory of Badr both an identity-based and historical dimension.

This analysis also highlights the geography of the battle (tag *placename*, 12%), describing strategic locations such as encampment areas, sites of confrontations, wells, and so forth. The toponymic study will enable tracking the mobility of the belligerents and, ultimately, comparing their frequency throughout the conflict.

Furthermore, the typology of objects (tag *object*, 9.9%), whether weapons, clothing, relics, or other symbols, is refined by categories. Additionally, with the recent restructuring of the database, the tags *animal* (2.7%) and *term* (10.9%) will be integrated into the tag *rs* (5.5%) to enrich it. This will enable the centralisation and isolation of a specific vocabulary, paving the way for the construction of a taxonomy of roles, titles, abstract concepts, or non-human entities.

The Database

The extracted data are stored within a relational (SQL) database that allows for their storage and structuring. This database is managed via the phpMyAdmin web application, which handles database administration (MySQL). This application enables the visualisation and verification of data, the debugging of queries, and direct interaction with the sixteen tables contained within the database.

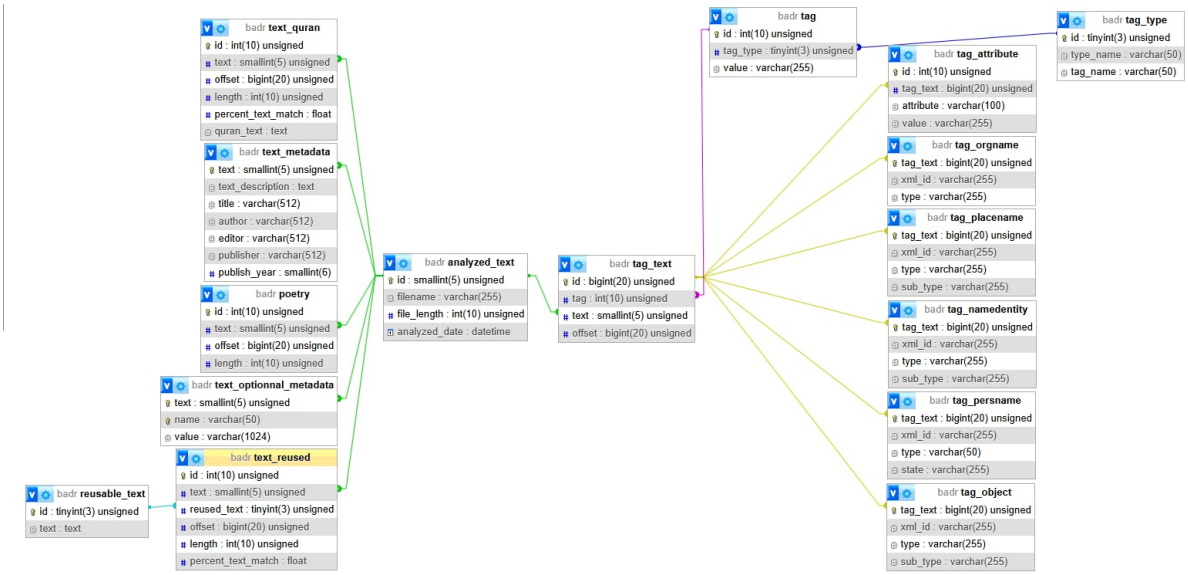


Fig 9. The Badr project database on phpMyAdmin

These sixteen tables store structured data as follows:

- **analyzed_text**: analysis data concerning text length, analysis date, and file name
- **text_metadata**: description of the text (title of the work, author's name, editor, publishing house, and year of publication, etc.)
- **text_optional_metadata**: optional metadata specific to each text
- **tag_text**: list of tags for a text with their position and value in the source text
- **tag**: tags mentioned in the guide without repetitions
- **tag_type**: tags referring to the encoding guide
- **tag_orgName**, **tag_placeName**, **tag_persName**, **tag_namedentity**, **tag_object**: specific data for each respective tag
- **tag_attribute**: occasional attributes
- **text_quran**: Quranic passages, storing their positions and lengths within the text
- **reusable_text**: list of patterns to be identified in the text
- **text_reused**: patterns found in the texts, storing their positions and lengths within the texts
- **poetry**: poetry within the texts

Research possibilities enabled by TEI-XML in the study of premodern Islamic texts

This paper concludes by outlining several research possibilities enabled by the proposed encoding guidelines and workflow in the study of premodern Islamic texts. These possibilities operate at three levels: macro-level analysis, such as the extraction and visualization of named entities across corpora ranging from a few thousand to tens of thousands of words; micro-level analysis, enabling the identification of narrative motifs and structural patterns embedded within the encoded texts; and historiographical analysis at a scale and granularity distinct from existing initiatives such as the KITAB project and the PASSIM tool.

Macro analysis of named entities

Applying these encoding guidelines and workflow could be used to map intellectual networks within the premodern Islamic world. By using TEI to encode named entities, whether they are compilers of hadith (*muḥaddithūn*), historians (*muʾarikhūn*), jurists (*fuqahāʾ*), exegetes (*mufassirūn*), or theologians (*mutakallimūn*), it becomes possible to track how these figures interacted over time and across regions. This approach could contribute to the study of how ideas, schools of thought, and religious practices spread across the Islamic world, supported by quantitative data for specific corpora. It could also help reveal relationships or oppositions between figures in these various fields of premodern Islamic texts, while enabling the extraction and comparison of *isnād*-s related to specific texts on a broader scale.

Encoding the geographic locations of cities, regions, and political entities mentioned in texts can also help track the political and cultural influence of different areas. For instance, TEI could be used to map how cities across Arabia (Mecca, Medina, Ṣanʿāʾ), Iraq (Kūfa, Baṣra, Wāsiṭ, and Baghdad), the Persian world, Syria (Damascus, Aleppo, Beirut), Egypt (Cairo), and al-Andalus (Cordoba, Seville, and others) played roles in the transmission of knowledge and the spread of Islamic traditions. This approach could also help identify the hierarchy of cities based on their contributions to specific Islamic fields, such as which locations were central to the development and spread of jurisprudence (*fatwa*), or how regional variations in texts, narratives, and practices emerged over time.

Micro analysis: looking for narrative motifs in texts

Micro-level analysis focuses on smaller, more specific elements within the texts, such as narrative motifs, recurring themes, symbolic structures, or literary patterns. This type of analysis involves close reading to identify how particular motifs—whether religious, legal, historiographical, or literary—are embedded and reappear across different parts of a text or across multiple texts. The proposed encoding guidelines and workflow can support the systematic identification and tagging of such motifs using TEI, making it possible to trace their transformations over time. This approach is not only relevant for literary studies, but also for exploring questions of memory, such as how a particular event is recalled, reinterpreted, or recontextualized in various sources. TEI further allows for the encoding of intertextual references and allusions, shedding light on how later authors borrow, adapt, or respond to earlier textual traditions. The goal here is not primarily quantitative analysis, but rather the qualitative identification and tracking of these traces, to better understand the dynamics of reuse, continuity, and innovation in premodern Islamic writing.

Better understanding of the evolution of premodern Islamic historiography

Analyzing the evolution of historiography through TEI encoding can allow for a better understanding of how historical narratives were constructed, what sources were considered authoritative, and how different scholars approached the task of recording and interpreting the past. By encoding multiple versions of historical works, TEI can allow to compare how different historians, such as al-Ṭabarī (d 310/923), al-Masʿūdī (d. 345/956), or even Ibn Khaldūn (d. 808/1406) approached historical writing in terms of their sources, methods, and political opinions.

Conclusions

This article contributes to the ongoing discussion on the role and benefits of text encoding within the field of digital humanities applied to Arabic literature. It presentes an encoding guidelines along with an innovative method for applying the TEI-XML standard to a corpus of classical Arabic texts. The study assesses the relevance and effectiveness of such encoding in the digital processing of the Badr project corpus, which comprises around forty texts dating from the early Islamic period to the 14th century. It highlights the advantages of using TEI-XML in terms of data structuring, interoperability, and textual readability. This encoding guidelines is an essential tool for the semi-automated or automated processing of textual data. It provides a rigorous working framework by establishing TEI-XML standards specifically adapted to the markup of texts in Arabic. This standardisation facilitates a systematic process of extracting, structuring, and storing named entities or narrative patterns in an SQL database. It thus paves the way for organised post-processing of the data, enabling meaningful visualisations and advanced analytical work. The workflow, whether simple or automated, for encoding Arabic texts was designed using exclusively open-source software such as VS Code and Notepad++. This approach deliberately avoids reliance on proprietary tools like *Oxygen XML Editor*, whose licensing costs can be prohibitive for individuals or institutions with limited funding — as is the case in our context. The aim is to ensure that the encoding process remains accessible, reproducible, and sustainable.

However, the proposed encoding guidelines and workflow also present certain limitations that should be acknowledged. As indicated in the title, the method is primarily designed for hadith-based texts structured around the association of *isnād* and *matn*. This format becomes progressively less dominant from the Mamluk period onward, particularly in the 13th

and 14th centuries. While the method is especially well-suited to genres such as hadith, historiography, jurisprudence, and exegetical works—where authors often explicitly trace the authority behind their statements—it may be less effective for texts that do not follow this structure. This includes certain philosophical treatises, *adab* literature, scientific works, or mystical and ascetic writings (*zuhd*), which may lack the formal *isnād* structure altogether. Additionally, the method cannot address the epistemological challenges posed by *isnād*-s themselves: many were constructed to project an illusion of authority, or are incomplete, abbreviated, or selectively formulated for various reasons. As such, while TEI encoding facilitates the systematic mapping of transmission chains and textual structures, it does not replace the need for critical textual analysis and historical judgment.

References

Brinis, Fazia, Mohammed Ourabah Soualah, and Farida Bouarab. 2023. "Comparative Study of Encoding Formats for Digitized Ancient Arabic Manuscripts." *Proceedings of the 24th International Arab Conference on Information Technology (ACIT)*, Ajman, United Arab Emirates.

Clivaz, Claire, and Sarah Schulthess. 2017. "Editing New Testament Arabic Manuscripts in a TEI-Base: Fostering Close Reading in Digital Humanities*." Journal of Data Mining & Digital Humanities*.

Cruse, Carl, and Sajawel Ahmed. 2024. "TafsirExtractor: Text Preprocessing Pipeline Preparing Classical Arabic Literature for Machine Learning Applications." *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, 67–73. @ LREC-COLING 2024.

Kimball, Ralph, and Joe Caserta. 2011. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. New York: Wiley.

Kozae, Mahmoud. 2022. "New Approaches to Tackle Textual Variation." In *An Unruly Classic: Kalīla and Dimna and Its Syriac, Arabic, and Early Persian Versions*, edited by Beatrice Gruendler and Isabel Toral-Niehoff, 211–222. Leiden: Brill.

Lancioni, Giuliano, and N. Peter Joosse. 2016. "The Arabic Diatessaron Project: Digitalizing, Encoding, Lemmatization." *Journal of Religion, Media and Digital Culture* 5 (1): 205–227.

Maraoui, Hajer, Kais Haddar, and Laurent Romary. 2017. "Modeling of al-Hadith al-Shareef with TEI." *Proceedings of the International Conference on Engineering & MIS (ICEMIS)*.

Maraoui, Hajer, Kais Haddar, and Laurent Romary. 2018. “Segmentation Tool for Hadith Corpus to Generate TEI Encoding.” In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, 252–260.

Maraoui, Hajer, and Kais Haddar. 2015. “74 automatisation de l’encodage des lexiques arabes en TEI.” *Deuxième colloque pour les étudiants chercheurs en traitement automatique du langage naturel et ses applications*, Sousse, Tunisia, 74–82.

Nahli, Ouafae, and Angelo Mario Del Grosso. 2020. “Creating Arabic Lexical Resources in TEI: A Schema for Discontinuous Morphology Encoding.” *6th IEEE Congress on Information Science and Technology (CiSt).*

Nahli, Ouafae, and Angelo Mario Del Grosso. 2021. “Structuring Arabic Lexical and Morphological Resources Using TEI: Theory and Practice.” *Research Challenges in Digitalization and Societal Transformation* 5 (3): 314.

Olivieri, Simona. 2018. “TEI-Encoding of Classical Arabic Grammatical Sources.” *DigiLex*, December 17.

Olivieri, Simona, Ivana Pepe, and Ilaria Cicola. 2016. “Encoding Arabic Rhetorical Structure: A Methodology for the Extraction of Arabic Lexical Information from TEI-Encoded Classical Sources.” *Proceedings of the 4th IEEE International Colloquium on Information Science and Technology (CiSt)*.

Salah, Mohammed Ourabah, and Mohamed Hassoun. 2012. “A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts.” In *Selected Papers from the TEI Conference*.

6. Biographical statements:

Adrien de Jarmy is a historian and islamologist specializing in the early Islamic period. Associate Professor at the University of Strasbourg, he earned his doctorate in 2023 at Sorbonne University with a thesis on the historiographical and normative constructions of the figures of the Prophet Muḥammad up to the 3rd/9th century. His work combines philology, digital humanities, and critical historiography. He has been a fellow of the IDEO and IFAO in Cairo and the DAAD in Germany, and is developing research on early Islamic texts using computational methods.

Clarck Junior Membourou Moimecheme is a historian of medieval Islam, specializing in the political and social history of Mecca under Mamluk rule. He earned his PhD in 2022 from the University of Brest, with a dissertation on the Banū Qatāda emirs of Mecca. He is currently a postdoctoral researcher in the BADR project at the University of Strasbourg. His work combines Arabic narrative sources, prosopography, and digital tools.

7. Project general information:

- Funder: Institut français d'islamologie (IFI): https://institut-islamologie.fr/.
- Institution: University of Strasbourg, France
- Principal investigator and team names and affiliations: Adrien de Jarmy (PI), Clarck Membourou Moimecheme, Renaud Soler, Eric Vallet (all University of Strasbourg)
- Duration: November 2022-November 2025
- Project official name: BADR
- Project type: research project
- Project URL: https://islamologie.unistra.fr/recherche/programmes-de-recherche/.