Armenian AutoEpiDoc: Automated Extraction and Encoding of Armenian Inscriptions into EpiDoc TEI/XML

Hamest Tamrazyan
DHLAB, EPFL, Lausanne, Switzerland
hamest.tamrazyan@epfl.ch

Emile Cornamusaz
EPFL, Lausanne, Switzerland
emile.cornamusaz@epfl.ch

Emanuela Boros
DHLAB, EPFL, Lausanne, Switzerland

Abstract

Armenian epigraphy is extensively documented in printed scholarly corpora, yet lacks machine-readable editions that support interoperability or computational analysis. In this paper, we present *Armenian AutoEpiDoc*, a system that automatically converts expert-verified Armenian inscription records into EpiDoc-compliant TEI/XML files. Operating on curated and domain-validated data, AutoEpiDoc maps Armenian-specific metadata to EpiDoc structures through rule-based templates and schema-aware validation. The workflow significantly reduces manual encoding effort and provides a scalable path toward producing digital editions and integrating Armenian inscriptions into international epigraphic infrastructures.

1 Introduction

Armenia possesses one of the richest and most continuous inscriptional traditions in the Near East, spanning religious, commemorative, funerary, legal, and architectural contexts for over a millennium. Yet despite this abundance, no comprehensive machine-readable corpus exists, and Armenian inscriptions remain largely absent from digital epigraphic infrastructures. Recent surveys highlight the scale and geographic dispersion of the material: more than 80% of Armenian inscriptions lie outside the borders of the modern Republic of Armenia, often in regions affected by conflict or infrastructural risk (Greenwood, 2014; Tamrazyan, 2023). Major corpora, such as the multi-volume *Divan Hay Vimagrut̹yan* (DHV), are authoritative but remain non-digital and unstructured, preventing the field from benefiting from interoperability, searchability, semantic modelling, and long-term preservation offered by standards such as EpiDoc[1].

Without controlled vocabularies, structured metadata, or machine-actionable corpora, Armenian inscriptions cannot be linked to infrastructures such as EAGLE[2] or FAIR Epigraphy (Bianchini, 2023)[3]. Dominant epigraphic ontologies are largely grounded in Greco-Roman traditions and do not reflect the cultural specificity or typological diversity of Armenian inscriptions. As noted in studies of digital cultural heritage (Liu et al., 2023), such mismatches can lead to conceptual loss when local traditions are forced into externally defined categories.

This limitation is not due to a lack of scholarship. Armenian epigraphy is extensively documented in catalogues, surveys, and monographs; rather, the challenge lies in the absence of scalable mechanisms for converting expert knowledge into standard-compliant digital editions. Manual EpiDoc encoding requires both technical expertise and familiarity with Armenian epigraphic conventions, making large-scale digitization impractical (Tamrazyan and Hovhannisyan, 2024; Tamrazyan et al., 2025).

TEI[4] provides a widely adopted XML framework for structured text encoding, with EpiDoc as its epigraphic application profile. Armenian inscriptions, however, remain largely absent from TEI/EpiDoc-based digital corpora due to the lack of scalable tools for transforming expert-curated catalogues into machine-readable editions.

To address this gap, we introduce *Armenian AutoEpiDoc*, the first system designed to automatically generate EpiDoc-compliant TEI/XML editions from expert-curated Armenian inscription records. Unlike OCR-based or text-mining approaches, AutoEpiDoc operates on inscriptions

[1] EpiDoc is a community-developed application of the TEI Guidelines for the digital encoding, publication, and interchange of ancient and medieval epigraphic and papyrological texts. https://epidoc.stoa.org/

[2] https://www.eagle-network.eu/

[3] https://inscriptiones.org/

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

Proceedings of SIGHUM 2026, pages 11–15
March 28-29, 2026 ©2026 Association for Computational Linguistics

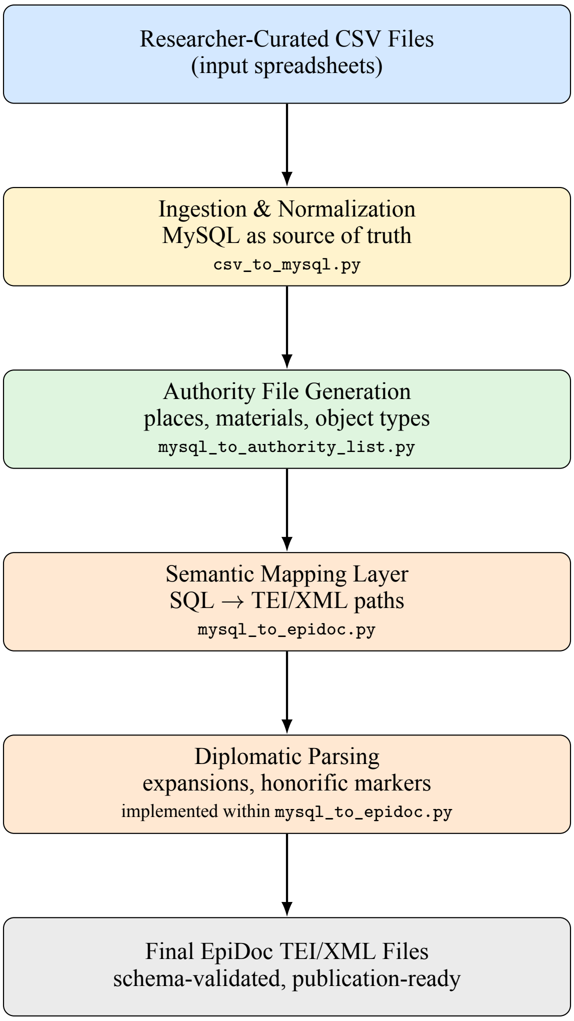
Researcher-Curated CSV Files
(input spreadsheets)
Ingestion & Normalization
MySQL as source of truth
csv_to_mysql.py
Authority File Generation
places, materials, object types
mysql_to_authority_list.py
Semantic Mapping Layer
SQL → TEI/XML paths
mysql_to_epidoc.py
Diplomatic Parsing
expansions, honorific markers
implemented within mysql_to_epidoc.py
Final EpiDoc TEI/XML Files
schema-validated, publication-ready

Figure 1: Color-coded architecture of the AutoEpiDoc ELT pipeline, showing Python modules and semantic processing layers.

that have already been validated by specialists and functions as a transformation layer that maps structured scholarly metadata and diplomatic transcriptions to EpiDoc elements, producing schema-valid XML. This significantly reduces the manual effort required for TEI encoding and enables the integration of Armenian epigraphic data into international digital epigraphic infrastructures and Linked Open Data (LOD) ecosystems.

In this paper, we present the design, implementation, and demonstration of *Armenian AutoEpiDoc*. All components of the system, including authority lists, generated EpiDoc files, and the complete AutoEpiDoc at `https://github.com/dhlab-epfl/autoepidoc`.

2 Methodology

Armenian AutoEpiDoc implements a semi-automated Extract–Load–Transform (ELT) pipeline designed to bridge the gap between unstructured research data and the strict hierarchical requirements of the EpiDoc standard. The architecture prioritizes data integrity and semantic interoperability through a three-stage process: (1) **Ingestion and Normalization** (Blue), (2) **Authority Management**, and (3) **XML Serialization**.

Figure 1 provides an overview of the AutoEpiDoc ELT pipeline. The color-coded layers in the figure correspond directly to these three stages:

Next, we clarify how data flows through the system and how each Python module operationalizes one component of the transformation pipeline into fully standards-compliant EpiDoc files.

2.1 Pipeline Architecture

The system utilizes an intermediate relational database (MySQL) as the primary source of truth rather than processing CSV files directly into XML. This decision allows for strict type enforcement and relational integrity checks prior to encoding.

2.1.1 Stage 1: Ingestion and Normalization

The input researcher-curated CSV files were organized into thematic datasets describing inscriptions, monuments, places, materials, and controlled vocabularies. Table 1 provides an overview of all input files and their roles in the ingestion workflow.

The ingestion module (`csv_to_mysql`) reads researcher-curated spreadsheets (*.csv*) as presented in Table 1 and it applies a heuristic normalization algorithm that sanitizes inconsistent column headers (e.g., handling variants such as *"sub-monument ID"* vs. *"sub_monument_id"*) and enforces type safety to avoid accidental numerical inference in inventory codes. The module then loads the cleaned data into a normalized Star Schema, where a central inscription table references auxiliary lookup tables for monuments, locations, materials, and object types.

2.1.2 Stage 2: Authority Management

A important requirement for modern digital epigraphy is the use of controlled vocabularies to guarantee semantic consistency. Instead of storing

CSV File	Description	Key Columns / Fields
`inscriptions.csv`	Primary dataset describing each inscription record; forms the core of the pipeline.	Inscription ID, monument ID, transcription, Armenian/English description, date (Armenian Era), material, object type, condition, bibliography.
`monuments.csv`	List of monuments and architectural structures to which inscriptions belong.	Monument ID, name (hy/en), type, village/-town, region, coordinates, parent monument ID (for complexes).
`locations.csv`	Hierarchical geographic information; used for authority file generation.	Place ID, name (hy/en), type (village, district, region), parent place ID, contemporary state name, coordinates.
`materials.csv`	Controlled list of inscription materials used to generate TEI `<material>` vocabularies.	Material ID, Armenian label, English label, notes, AAT reference (if available).
`object_types.csv`	List of inscription-bearing object types used across Armenian epigraphy.	Object type ID, Armenian label, English label, definition, AAT reference (if available).
`bibliography.csv`	Structured citations for printed corpora, surveys, and monographs.	Bib ID, author(s), year, title, publication, pages, stable ID (e.g., Zotero).
`photos.csv` (optional)	Links between inscriptions and photographic documentation.	Photo ID, inscription ID, file path/URL, photographer, date, rights.

Table 1: Researcher-curated CSV (spreadsheets) files ingested by the AutoEpiDoc pipeline and their role in the MySQL-based normalization stage. Each file populates one or more relational tables used as the authoritative source before XML serialization.

metadata as unstructured strings, AutoEpiDoc generates standalone TEI-compliant Authority Files for **Places**, **Monuments**, **Materials**, and **Object Types**.

The authority generation module (`mysql_to_authority_list`) builds these files hierarchically. For geospatial entities, relational foreign keys (e.g., *parent_place_id*) are converted into semantic TEI nodes (`<note type="relation">`), thereby reconstructing administrative hierarchies such as those of the Artsakh region. The module also integrates LOD principles by aligning local terminology with international ontologies. Materials and Object Types are automatically linked to the Getty Art & Architecture Thesaurus (AAT) and to EAGLE vocabularies via a dedicated `<standOff>` section, ensuring interoperability with European digital infrastructures.

2.1.3 Stage 3: XML Serialization

The final stage of the pipeline converts MySQL inscription records into EpiDoc-compliant TEI/XML files. The transformation script (`mysql_to_epidoc`) uses a dynamic XPath-based mapping strategy to translate flat SQL rows into nested TEI structures, including both the `<teiHeader>` and the `<text>` body.

Diplomatic Transcription Parsing Metadata and transcription are processed separately. Armenian epigraphic transcriptions rely on scholarly diplomatic conventions that are meaningful to specialists but not directly machine-readable, such as letters functioning as numerals (e.g. date formulas), non-punctuation symbols used as numeric separators, abbreviated forms, ligatures, and overlines marking honorific or sacred names.

Transcriptions are therefore stored in a simplified ASCII notation reflecting established Armenian epigraphic practice and are subsequently parsed by a rule-based module that maps these conventions to structured EpiDoc TEI/XML. For example, abbreviated forms are encoded using `<expan>` with explicit `<abbr>` and `<ex>` components, numeric letters are normalized using `<num>`, and overlines and ligatures are represented using `<hi>` and `<join>`. This preserves palaeographic and editorial information while making Armenian diplomatic transcriptions interoperable with international digital epigraphic standards, without requiring researchers to write TEI/XML manually.

The most challenging cases for rule-based parsing involve damaged numeral sequences, ambiguous ligatures, and inconsistent scholarly transcription practices across sources. In such cases, AutoEpiDoc preserves the original transcription and flags uncertain elements for expert review rather than enforcing potentially erroneous normalization.

Chronological and Bilingual Processing To support the needs of Armenian studies, the system represents dates in both the Armenian Era and the Gregorian calendar. Original Armenian-era dating expressions are encoded in `<origDate>`

with a custom @calendar attribute, while computed Gregorian equivalents are provided for indexing and interoperability. All descriptive metadata is generated bilingually (Armenian/English) using xml:lang attributes to support multilingual search and display.

3 Output Examples

In this section, we present some excerpts from the EpiDoc-compliant TEI/XML files generated by AutoEpiDoc from real Armenian inscription records obtained in this work.

Example 1: TEI Header and Authority Integration. The <teiHeader> generated by AutoEpiDoc integrates descriptive and administrative metadata derived from the underlying relational schema and enriched through automatically generated authority lists. Stable URN-based identifiers for monuments, architectural components, materials, object types, techniques, and scripts are referenced via @ref attributes, while multilingual naming is supported through parallel language-specific elements. An inscription associated with the Gandzasar monastic complex is encoded as follows:

```
<msDesc xml:id="
    ms_Gandzasar_Monastic_complex__monastery">
  <msIdentifier>
    <repository ref="urn:armepic:mon:MON0002">
      <objectName xml:lang="en">Gandzasar
          Monastic complex / monastery</
          objectName>
      <objectName xml:lang="hy">Գանձասար
          վանական համալիր / վանք</objectName>
    </repository>
  </msIdentifier>
  <msPart xml:id="ms_Gavit_Narthex">
    <msIdentifier>
      <repository ref="urn:armepic:mon:
          MONPART0010">
        <objectName xml:lang="en">Gavit (
            Narthex)</objectName>
        <objectName xml:lang="hy">Գավիթ</
            objectName>
      </repository>
    </msIdentifier>
  </msPart>
  <physDesc>
    <objectDesc>
      <supportDesc>
        <support>
          <objectType ref="urn:armepic:
              objecttype:OBJ0006" xml:lang="en"
              >lintel</objectType>
          <objectType xml:lang="hy">բարավոր</
              objectType>
          <material ref="urn:armepic:material:
              MAT0001" xml:lang="en">tuff</
              material>
          <material xml:lang="hy">տուֆ</
              material>
          <rs type="technique" ref="urn:armepic:
              technique:TEC001" xml:lang="en">
              carved</rs>
          <rs type="technique" xml:lang="hy">
              փորագիր</rs>
        </support>
      </supportDesc>
      <layoutDesc>
        <layout xml:lang="hy">Հյուսիսային մուտքի
            կիսակամար քարին,
            արտաքուստ</layout>
        <layout xml:lang="eng">On the stone
            semi-arch of the northern entrance,
             from the outside</layout>
      </layoutDesc>
    </objectDesc>
    <handDesc>
      <handNote xml:id="hand_ART0001" scriptRef
          ="urn:armepic:script:SCR002">
        <term xml:lang="en">Bolorgits Erkatagir
            </term>
        <term xml:lang="hy">Բոլորգիծ
            երկաթագիր</term>
      </handNote>
    </handDesc>
  </physDesc>
</msDesc>
```

Listing 1: Example of Armenian epigraphic encoding (TEI/XML)

This example illustrates the transformation of catalog descriptions into semantically structured metadata aligned with international vocabularies.

Example 2: Diplomatic Transcription Encoding. AutoEpiDoc converts Armenian diplomatic transcription conventions into structured TEI markup while preserving palaeographic and editorial information. Line breaks, ligatures, abbreviations, expansions, and numerals are encoded explicitly, supporting both faithful representation and computational reuse.

```
<text>
  <body>
    <div type="edition">
      <ab>
        <lb n="1"/><w>։Թիւ։</w>
        <w>Չհ
          <num value="1271">1271</num>
        </w>
        <lb n="2"/>
        <w>Կ
          <hi rend="ligature">ամ</hi>
          <hi rend="ligature">աւ</hi>
          <hi rend="ligature" xml:id="lig1">Ն</
              hi>
        </w>
        <w>
          <hi rend="ligature" xml:id="lig2">Ա</
              hi>
          <emph rend="bold">
            <expan>
              <abbr/>
              <ex>ստուծո</ex>
              <abbr>յ,</abbr>
            </expan>
          </emph>
        </w>
        <join xml:id="j1" result="ligature"
             target="#lig1 #lig2"/>
        <w>ես`</w><w>Յոհանես,</w>
      </ab>
    </div>
  </body>
</text>
```

Listing 2: TEI/XML encoding of an Armenian inscription edition

Example 3: Chronological Representation. Dates in Armenian epigraphic corpora are often expressed using the Armenian Era (Ա.Թ.). AutoEpiDoc encodes both the original dating expression and a normalized Gregorian equivalent, enabling chronological interoperability while preserving historical conventions.

```
<origin>
  <origPlace>
    <!-- place information -->
  </origPlace>
  <origDate>
    <date calendar="#cal_armenian" when-
        armenian="0720">Չի</date>
    <date calendar="#cal_gregorian" when="1271
        ">AD 1271</date>
  </origDate>
</origin>
```

Listing 3: Encoding of original date using Armenian and Gregorian calendars

4 Conclusions

The examples presented in this paper show that AutoEpiDoc is not merely a tool for converting tabular data into XML, but a system for semantically restructuring Armenian inscription records. Rather than reproducing the flat structure of the source data, the pipeline generates hierarchical TEI/EpiDoc documents that preserve diplomatic fidelity while supporting long-term digital preservation, interoperability, and reuse.

The resulting outputs integrate multilingual metadata, Armenian-era dates with normalized Gregorian equivalents, controlled vocabularies, and explicit authority references in a fully machine-actionable form. Schema-based validation and rule-driven transformations ensure consistency and scalability across large and heterogeneous corpora.

AutoEpiDoc has been applied to a growing subset of Armenian inscriptions within the ArmEpiC corpus and the Artsakh Epigraphy Atlas. In pilot comparisons, manual TEI/EpiDoc encoding typically required tens of minutes per inscription, whereas the system produces a first-pass EpiDoc file in seconds, requiring only targeted expert correction. Domain specialists confirmed that the generated encodings preserve Armenian epigraphic conventions while substantially reducing manual encoding effort. The system is currently in active use and continues to be iteratively validated within these projects, demonstrating its value for real epigraphic research workflows. Remaining challenges concern ambiguous ligatures and damaged numeral sequences, which are flagged for expert review.

By aligning Armenian epigraphic data with international TEI/EpiDoc standards, this work lays the foundation for the first comprehensive, standards-aligned digital corpus of Armenian inscriptions and enables interoperability with epigraphic infrastructures such as FAIR Epigraphy, EFES, and Linked Open Data environments.

References

Francesco Bianchini. 2023. Looking beyond the text: Opportunities and challenges in the digitisation of sanskrit inscriptions. *Can't Touch This*.

Tim Greenwood. 2014. Armenian epigraphy. In *Armenian Philology in the Modern Era*, pages 101–121. Brill.

Fangchao Liu, John Hindmarch, and Mona Hess. 2023. A review of the cultural heritage linked open data ontologies and models.

Hamest Tamrazyan. 2023. Digitization of the inscriptions on the monuments of armenian cultural heritage in nagorno-karabakh region. In *DH*.

Hamest Tamrazyan and Gayane Hovhannisyan. 2024. Digital guardianship: Innovative strategies in preserving armenian's epigraphic legacy. *Heritage*, 7(5):2296–2312.

Hamest Tamrazyan, Gayane Hovhannisyan, and Arsen Harutyunyan. 2025. From stone to standards: A digital heritage interoperability model for armenian epigraphy within the leiden and epidoc frameworks.