École Polytechnique Fédérale de Lausanne

AutoEpiDoc: Automated Encoding of Armenian Inscriptions into EpiDoc format

Emile Cornamusaz

Project Supervisors:
Dr. Hamest Tamrazyan
Dr. Emanuela Boros

7th January 2026

Abstract

This report presents the development of *AutoEpiDoc*, a semi-automated data processing pipeline designed for the digital encoding of Armenian epigraphic inscriptions into the international EpiDoc (TEI-XML) standard. Armenian cultural heritage possesses a historically significant corpus of inscriptions, but much of this data remains stored in fragmented, non-standardized tabular formats, limiting its interoperability and long-term preservation. The methodology employs a modular architecture using Python-based scripts to facilitate the transition from raw researcher-curated spreadsheets to a structured relational database (MySQL). This intermediate database acts as a source for the automated generation of two outputs: (1) standardized, hierarchical authority files for controlled vocabularies, such as materials, places, and monuments; and (2) individual, machine-readable EpiDoc XML files for each inscription.

A central contribution of this work is the implementation of a hybrid workflow that addresses the specific complexities of Armenian epigraphy, including bilingual metadata handling. The system also integrates a specialized diplomatic notation parser to transform ASCII-based transcriptions into valid TEI-XML edition blocks.

The project represents the *first tentative* effort to systematically digitize an Armenian epigraphic corpus, establishing a reproducible framework for future digital humanities initiatives. While the pipeline automates most of the structural and metadata encoding, the workflow incorporates a critical scholarly threshold for manual curation, allowing researchers to refine complex features such as lemmatization and paleographic punctuation. Ultimately, *AutoEpiDoc* provides a scalable solution for the large-scale digitization of Armenian heritage, ensuring its accessibility and longevity in the global digital landscape.

Contents

Chapter 1

Introduction

1.1 Context

The digitization and preservation of cultural heritage gained increasing importance in digital humanities. Epigraphic inscriptions (i.e. texts carved on stones, monuments or architectural elements) represent an important source of historical and cultural information. However, these data are contained in different data formats and archives, thus missing standardization and interoperability.

To address these challenges, the ”AutoEpiDoc” project aims to create a unified digital corpus of inscriptions from the Armenian cultural heritage. The project uses EpiDoc, an international XML standard based on TEI, widely adopted for encoding ancient texts and inscriptions in a structured and interoperable format.

In this context, the work presented here focuses on automating the transformation of tabular data stored in spreadsheets or CSV files into a structured database and then EpiDoc XML files that follow the metadata and encoding conventions. This process supports the long-term goal of integrating Armenian epigraphic data into accessible digital infrastructure.

1.2 Motivation

The motivation behind this work is to bridge the gap between raw research data and standardized digital representation in Armenian epigraphy. Before this project, inscriptions and associated metadata were stored accross multiple spreadsheets, making data management and validation difficult. Manual encoding of each inscription into EpiDoc XML was time-consuming, error-prone and not scalable for a large corpus.

By introducing a semi-automated pipeline, this work reduces the time and effort required to generate valid records. As humans are prone to errors and inconsistency, the conversion also minimizes the error risk and ensures consistency between inscriptions since the outputs all derive from a single relational database. The adherence to EpiDoc standards also facilitates interoperability with international databases such as EAGLE. Ultimately, this project represents a concrete step toward the creation of a sustainable and reproducible digital workflow for Armenian epigraphic

data, laying the baseline for future research and collaboration in the digital humanities field.

1.3 Challenges

The most significant challenge was the project's status as the **first tentative** attempt to digitize a corpus of Armenian epigraphy. Unlike Latin or Greek epigraphy, Armenian epigraphy lacked a standardized machine-readable methodology.

- **Absence of Precedents:** As a pioneering effort, the project could not rely on existing "best practices" for Armenian-specific features. We had to architect the initial mapping between traditional notation and TEI-XML from the ground up.
- **Scaling the "First Tentative":** Because this script serves as the baseline for the entire Armenian Epigraphic Corpus (ArmEpiC), every technical choice regarding URN structures and metadata labels had to be designed for long-term scalability and interoperability.

1.4 Goals

The goal of this work is to design and implement a data processing pipeline capable of transforming raw epigraphic data from CSV format into standardized EpiDoc XML files through an intermediate MySQL relational database.
This pipeline ensures that the epigraphic records can be easily validated, queried and exported in a format suitable for publication, preservation and interoperability.

Concretely, the objectives of this project are :

1. **Data integration:** Develop a script that imports and normalizes tabular data from multiple spreadsheets into a local MySQL database, maintaining relational links between entities through their unique IDs.
2. **Format alignment:** Map database fields to the corresponding elements in the EpiDoc XML model, ensuring consistency with the project's metadata requirements.
3. **Automated metadata generation:** Develop a script capable of querying the database and producing one XML file per inscription, embedding all relevant contextual information (monument, material, script, technique, etc...) into the metadata fields of the EpiDoc output.
4. **Automated text formatting:** Develop a script that transforms the raw text inscriptions into xml encoding following the EpiDoc format requirements.
5. **Authority File Generation:** Design and implement a module to generate standalone, TEI-compliant authority lists (for materials, bibliographies, places, monuments, etc.) from the database.
6. **Reusability:** Ensure that the developed tools can be reused and adapted for future datasets by properly commenting and documenting the code.

1.5 Code and Data Availability

The complete source code for the *AutoEpiDoc* pipeline, including the `csv_to_mysql.py`, `mysql_to_authority_list_updater.py`, and `mysql_to_epidoc.py` scripts, is hosted on GitHub. The repository also contains sample datasets and the required data format. It is meant as as an open-source repository to encourage community contributions and long-term sustainability.

`https://github.com/dhlab-epfl/autoepidoc`

Chapter 2

Methodology

2.1 System Architecture and Workflow

The core methodology of the ”AutoEpiDoc” project relies on a semi-automated data pipeline that extracts, loads and transforms the raw epigraphic data. This architecture was chosen to control the transition from semi-structured researcher data (CSV spreadsheets) to highly structured interoperable XML data (EpiDoc).
The workflow consists of three distinct stages, each handled by a dedicated python module :

1. **Ingestion & Normalization:** Raw CVS data is ingested, cleaned and normalized into a local Relational Database management system (MySQL).
2. **Inscription Encoding:** Individual inscriptions records are queried from the database and wrapped in EpiDoc-compliant XML structures.
3. **Authority Management:** Controlled vocabularies are generated as standalone TEI XML files to serve as the semantic backbone of the corpus.

This modular approach ensures that the database is the single source of information and prevents synchronization errors that can occur when manually editing XML files.

2.2 Data Ingestion and Database design

2.2.1 Data Ingestion Strategy

The first stage of the pipeline addresses the variety of input data. The script `csv_to_mysql.py` uses the *pandas* library to read CSV inputs as strings to preserve data integrity before continuing. To ensure the data are compatible with MySQL column naming conventions, a cleaning function is applied to the raw CSV headers. This function strips whitespaces, replaces special characters with underscores and forces lowercase.
The script connects to the local MySQL host using the *SQLAlchemy* library. It automatically checks for the existence of the target database and creates it if it does not exist. The input directory is then iterated through, treating each CSV file as a separate database table. To make

iterative testing easier, the script replaces already existing databases so that the tables can be fully rebuilt from the source CSVs when the schema changes.

2.2.2 Relational Database Schema

The Database schema is designed to separate the inscription text from its metadata, using a star-schema structure where the central inscription table references surrounding tables. The database contains the following entities :

- *epigraphysamples:* The central table containing the inscription text and foreign keys to metadata.
- *listplaces:* Contains geographical data, including coordinates and parent-place relationships.
- *listmonum & listsubmonum:* Hierarchical tables defining the monuments (e.g., churches, monasteries) where inscriptions are found.
- *listmat:* Defines physical materials (e.g., Tuff, Basalt) and links them to external vocabularies like AAT or EAGLE.
- *listobjs:* Categorizes the support carrier (e.g., Khachkar, Wall).
- *listtechniques:* Describes the method of inscription (e.g., engraved, relief).
- *listscripts:* Defines the paleographic script style (e.g., Erkatagir, Bolorgir).
- *listbibl*: A bibliography table for literature references.

2.3 Automated Inscription Encoding

The most important stage of the pipeline, handled by `mysql_to_epidoc.py`, aggregates data from the central *epigraphysamples* table and the various lookup tables to produce one XML file per inscription.

2.3.1 TEI Header Construction

The script dynamically constructs the `<teiHeader>`, which contains the metadata description of the inscription.

- File Description (fileDesc): The title and publication statements are generated bilingually (English and Armenian). The script automatically assigns a Unique Resource Name (URN) following the format urn:armepic:artsakh:ins:inscription_id.
- Source Description (sourceDesc): This section details the physical object. The script executes SQL joins to retrieve the preferred English and Armenian labels for the monument, sub-monument, material, and object type. These are inserted into the XML with ref next pointing to the Authority Lists that will be generated in the next step (e.g., ref=”urn:armepic:mon:monument_id”).

- History and Provenance: The `<history>` element is populated with the place of origin and the date. The script distinguishes between the `place_find` (provenance of discovery) and `place_geo` (original location), mapping them to <provenance> and <origPlace> respectively. It also handles multiple dating systems, encoding dates for both the Armenian Era and the Gregorian calendar.

2.3.2 Text Handling

Finally, the inscription text itself is processed. The script imports a helper function `dhv_to_epidoc` to transform the transcribed text into TEI-compliant XML. This is wrapped in a `<div>` of type `edition`, while descriptive fields from the database are mapped to `<div>` elements of type `commentary` and `bibliography`.

2.3.3 Validation and Serialization

Before writing the file, the XML tree is converted to a string and ”prettified” for human readability using the *minidom* library. A regex cleaning step removes the default XML declaration and replaces it with the specific xml-model processing instructions required for EpiDoc validation (referencing the RelaxNG schema). This ensures that the output is not just well-formed XML, but valid EpiDoc.

2.4 Generation of Controlled Vocabularies (Authority Lists)

A critical requirement for EpiDoc compliance is the use of controlled vocabularies. Instead of embedding free text, the project generates standalone TEI authority files. The script *mysql_to_authority_list.py* handles this process.

2.4.1 Iterative XML Updates

Unlike the inscription generation, which creates new files, the authority list updater is designed to respect existing data. The script checks if a target XML file (e.g., `ArmEpiC_ListPlace.xml`) already exists. If it does, it parses the existing tree to preserve the revision history before updating or adding new entries. If the file is missing, it creates it, generating a new TEI root with a standard header defining the publication statement, licensing, and project details.

2.4.2 Semantic Mapping and External Linking

The script maps MySQL rows to XML structures. A notable feature of this module is the integration of external references. The script pulls external URIs from the database—such as EAGLE, Wikidata, and Pleiades—and encodes them into the XML.

Chapter 3

Implementation Details

3.1 Data Ingestion and Normalization

The first challenge in the pipeline is to bridge the gap between user-provided CSV data and strict schema requirements of a relational database. The ingestion process is handled by the script `csv_to_mysql.py`.

3.1.1 Cleaning and standardization

Data entered by researchers often includes inconsistencies in column names (e.g., `Place ID` vs `place_id` vs `Place-ID`). To tackle this, the script implements a normalization function `clean_colname` that processes raw names before table generation. The code performs the following transformations:

1. **Stripping:** Removes leading and trailing whitespaces.
2. **Normalizing:** Replaces spaces, hyphens and slashes with underscores.
3. **Filtering:** Removes non-alphanumeric characters (except underscores) to ensure valid SQL IDs.

```
    colname = ''.join(ch for ch in c if ch.isalnum() or ch == '_')
```

4. **Case Folding:** Converts all headers to lowercase.
5. **Default Case:** If the resulting string is empty, it defaults to `col` to prevent SQL errors.

For example, a CSV column named "*sub-monument ID*" is automatically converted to "*sub_monument_id*" in the database.

3.1.2 Type Safety during Ingestion

An important technical constraint was to ensure the *pandas* library doesn't infer data types. Inventory numbers (e.g., "001") and dates (e.g., "1271") can be misinterpreted as integers, causing the loss of leading zeros or formatting. To prevent this, the `read_csv` function is called

with `dtype=str` parameter so that every value is considered as a string. Additionally, a post-processing step normalizes empty strings to `pd.NA` using a regex replacement `r'^\s*$'` to ensure NULL values are stored in MySQL instead of empty strings.

3.1.3 Database Connectivity and Optimization

The data ingestion script does not assume the existence of the target database. To ensure robustness, the script first establishes a temporary connection to the MySQL server root. It then executes a `CREATE DATABASE IF NOT EXISTS` command with `utf8mb4` character set encoding. This is critical for supporting the specific Armenian Unicode characters found in the dataset.

For performance optimization during the upload of large datasets, the script utilizes SQLAlchemy's execution options. The `to_sql` method is configured with:

- `if_exists='replace'`: Allows for iterative development by rebuilding the schema on every run.
- `method='multi'`: Passes multiple rows in a single `INSERT` statement.
- `chunksize=1000`: Batches records to prevent memory overflow during the transaction.

3.2 Inscription XML Construction

3.2.1 Dynamic XPath Mapping

The `mysql_to_epidoc.py` script creates the XML file using a direct mapping between SQLAlchemy rows and XML ElementTree nodes. The construction of the `<sourceDesc>` uses conditional logic to handle missing data. For example, the repository (monument) mapping performs a secondary SQL lookup to fetch the monument's name:

```
mon = conn.execute(sql_text("SELECT preferred_name_eng FROM listmonum WHERE
    auto_id = :mid"), ...)
# ...
if mon and mon.get("preferred_name_eng"):
    ET.SubElement(repository, "objectName", attrib={"xml:lang": "en"}).text
    = mon["preferred_name_eng"]
```

This ensures that missing values will not prevent the script from executing or alter the integrity of the data.

3.2.2 Chronological and Calendrical Processing

The dataset distinguishes between the visual date carved on the stone (often using the Armenian numeral system) and the computed Gregorian date. The `mysql_to_epidoc.py` script handles this by generating multiple `<origDate>` elements with distinct `@calendar` attributes.

The script checks for the existence of both the Armenian and english date strings and formats them it with specific tags:

```
ET.SubElement(origDate, "date", attrib={"calendar": "#cal_armenian", "when-
    armenian": year_part}).text = text_part
```

```
...
ET.SubElement(origDate, "date", attrib={"calendar": "#cal_gregorian", "when
    ": parsed_date}).text = record["date_display_en"]
```

This preservation ensures that the digital record remains faithful to the paleographic reality of the artifact while still providing a machine-readable Gregorian date for indexing.

3.2.3 Parsing the Transcription (Helper Integration)

The raw transcription text is stored in the `Text_T` column using a simplified diplomatic notation. The implementation imports a custom helper `dhv_to_epidoc` to process this string. Instead of manual string replacement, this modular function parses the text and returns a structured XML object, which is then appended to the `div[@type='edition']` node.
The diplomatic notation is transformed into EpiDoc tags using regex expression, for example for gaps of unknown length :

```
# [---] -> <gap reason="lost" extent="unknown" unit="character"/>
text = re.sub(r'\[---\]', r'<gap reason="lost" extent="unknown" unit="
    character"/>', text)
```

3.2.4 Bibliography

The system handles bibliographic references not just as text, but as linked entities. The source column `bibliography` contains a semicolon-separated list of IDs (e.g., `BIBL_007; BIBL_002`).

The script parses this string to populate the biblography section of the XML file. It generates a `<listBibl>` in the text body, resolving the ID to the full citation string (Author, Year, Title) via a secondary SQL query.

```
list_bibls = [b.strip() for b in record["bibliography"].split(";") if b.
    strip()]
n = 1
for bibl in list_bibls:
    # Query type for facsimile mapping
    bibltype = conn.execute(sql_text("SELECT type FROM listbibl..."))
    ET.SubElement(surface, "graphic", attrib={
        "n": str(n),
        "source": f"urn:armepic:bibl:{bibl}",
        "ana": bibltype["type"]
    })
    n += 1
```

3.2.5 Post-Processing and Serialization

The standard Python *xml.etree.ElementTree* library does not support the specific processing instructions required by EpiDoc. To solve this, the main function performs a regex substitution on the final byte string:

1. **Minidom Prettify:** The tree is converted to a string with indentation.
2. **Header Injection:** The default `<?xml ...?>` declaration is stripped and replaced with the `<?xml-model ...?>` RelaxNG references.

3. **Whitespace Cleanup**: A regex is applied:

```
    re.sub(r'<lb([^>]*)/>\s*\n\s*', r'<lb\1/>', xml_text)
```

   to prevent the pretty-printer from breaking lines inside the transcription block, which would alter the semantic meaning of the diplomatic text.

3.3 Authority File Generation

3.3.1 The ”Update Safe” Parsing Strategy

The `mysql_to_authority_list.py` script differs from standard exporters by using a ”read-modify-write” model. The function `get_or_create_tei_root` attempts to parse the target XML file first. If the file exists, it extracts the existing `tei:revisionDesc` block to preserve the file's history before appending a new `change` element with the current timestamp and agent `urn:armepic:agent:updater_script`.

3.3.2 Building Hierarchical XML from Tables

SQL tables store hierarchy via foreign keys (e.g., `parent_place_id`), but TEI XML represents this via semantic linking. For example, the `update_place_auth` function programmatically converts the SQL foreign key into a TEI `<note>` element:

```
if row.get("parent_place_id"):
    ET.SubElement(plc, "note", attrib={
        "type": "relation",
        "target": f"urn:armepic:artsakh:plc:{row.get('parent_place_id')}"
    }).text = "partOf"
```

This specific snippet ensures that the generated XML graph accurately reflects the ”part of” relationship between a village and its district.

3.3.3 Handling Linked Open Data (LOD)

For material and object types, the script constructs a `<standOff>` section separate from the main body. The logic iterates through the input rows, detecting columns named `exactmatch` or `closematch`. It then aggregates these into a list of tuples (`xmlid`, `target_uri`, `matchtype`) which are rendered into a `listRelation` block. This separates the internal vocabulary definition from its external semantic alignment. A line containing LOD information in the final XML looks like:

```
<relation name="closeMatch" active="urn:armepic:object:OBJ0002" passive="
    https://www.eagle-network.eu/voc/objtyp/lod/416" />
```

3.3.4 Namespace Handling and XPath Context

To successfully update existing TEI files, the script must navigate the XML tree using namespaced XPaths. The `xml.etree.ElementTree` library requires explicit namespace registration. The script defines a global namespace dictionary to resolve the TEI URL (*http://www.tei-c.org/ns/1.0*).

```
NS = {
    'tei': 'http://www.tei-c.org/ns/1.0',
    'xml': 'http://www.w3.org/XML/1998/namespace'
}
# usage in search
header = root.find(".//tei:teiHeader", NS)
```

Without this explicit mapping, the script would fail to locate elements such as `tei:revisionDesc` or `tei:listPlace`, leading to file corruption or duplication of root elements.

Chapter 4

Results and Validation

4.1 Overview of Generated Corpus

The execution of the `mysql_to_epidoc.py` pipeline resulted in the successful generation of a structured digital corpus. The system processed the entirety of the relational database, producing two distinct categories of XML outputs:

- **Authority Files:** A set of hierarchical controlled vocabularies (Places, Monuments, Materials, Scripts) defining the semantic scope of the corpus.
- **Inscription Records:** Individual TEI-XML files for each epigraphic record, fully linked to the authority files via URNs.

The transformation process demonstrated high efficiency, converting complex relational data into standardized semantic XML in a single batch operation.

4.2 Case Study: Inscription ART0001

To validate the fidelity of the transformation, we analyze the processing of a specific record: *ART0001* (Gandzasar Monastery). This section compares the raw input stored in the MySQL `epigraphysamples` table against the final EpiDoc XML output. The full output file textit-ART0001.xml can be found in Appendix A.

4.2.1 Input Data Structure

The source data for record *ART0001* contains a mix of direct textual data, foreign key references, and specific Armenian formatting. Table 4.1 highlights the relevant raw values extracted from the *epigraphysamples* table before processing.

Table 4.1
Raw Database Entry for Inscription ART0001 in epigraphysamples table

Column Name	Raw Value
`inscription_id`	ART0001

Column Name	Raw Value
monument_id	MON0002
sub_monument_id	MONPART0010
place_origin	MON0002
place_find	MON0002
place_geo	PLC0009
date_display_arm	ՉԻ - 720
date_display_en	AD 1271
script_id	SCR002
material_id	MAT0001
technique_id	TEC001
object_id	OBJ0006
inscription_type	INSCT0001
condition_hy	PS027
description_hy	Փորագրությունը համաչափ է, արված վարպետ գրչի ձեռքով:
description_en	The carving is symmetrical and executed by a master scribe.
layout_text_hy	Հյուսիսային մուտքի կիսակամար քարին, արտաքուստ
layout_text_en	On the stone semi-arch of the northern entrance, from the outside
text_t (Snippet)	:Թիւ: ՉԻ (1271)/ Կ{ամ}{աւ}{ն ^^Ա}(ստուծո)յ^^, ես՝ Յոհաննէս...
Bibliography	BIBL_002; BIBL_003; BIBL_004; BIBL_006; BIBL_005; BIBL_001

4.2.2 Metadata Enrichment and Linking

One of the primary goals of the pipeline was to resolve opaque database IDs into human-readable, semantically linked metadata. In the generated `ART0001.xml` file, the script successfully resolved the foreign keys.

1. Monument Resolution: The input `MON0002` was resolved against the `listmonum` authority table. The output XML contains the full English title ”Gandzasar Monastic complex” and establishes a semantic link via the `ref` attribute.

```
<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>
```

2. Material and Technique: Similarly, the material code `MAT0001` and technique `TEC001` were resolved. Note that the system automatically injected the English labels ”Tuff” and ”Engraved” while maintaining the URN references for machine readability.

```
<supportDesc>
  <support>
    <material ref="urn:armepic:material:MAT0001" xml:lang="en">Tuff</
   material>
    <rs type="technique" ref="urn:armepic:technique:TEC001" xml:lang="en">
   Engraved</rs>
  </support>
</supportDesc>
```

4.2.3 Textual parsing and Apparatus

The transformation of the inscription text represents the most complex data manipulation. The helper function processed the diplomatic string, converting custom markers into valid TEI-XML tags. Here are some of the major convertions :

- **Line Breaks:** The forward slash / in the input was converted to `<lb n="..."/>`.
- **Expansion:** The braces `{...}` were parsed to separate the abbreviation from the expansion.

Generated Edition Block (Snippet):

```
<div type="edition">
    <ab>
        <lb n="1"/>
        <g type="punct" subtype="numeral-marker">:Թիւ:</g>
        ՉԻ
        <lb n="2"/>
        Կ
        <hi rend="ligature">ամ</hi>
        ....
        <expan>
            <abbr>Ա<abbr/>
            <ex>ստուծո</ex>
            <abbr>յ</abbr>
        </expan>
        ...
    </ab>
</div>
```

Note: The visualization above demonstrates the structural mapping of the symbols into TEI tags.

4.3 Schema Validation

To ensure interoperability with the European cultural heritage infrastructure, the generated files were validated against the official EpiDoc RelaxNG schema (`tei-epidoc.rng`).

The output files include the required `xml-model` processing instructions injected during the serialization phase:

```
<?xml-model href="https://www.stoa.org/epidoc/schema/9.7/tei-epidoc.rng"
  type="application/xml"
  schematypens="http://relaxng.org/ns/structure/1.0"?>
```

Validation tests confirmed that the structural logic—specifically the nesting of `msDesc`, the placement of `physDesc`, and the formatting of the `teiHeader`—complies with version 9.7 of the guidelines. This compliance guarantees that the corpus can be indexed by international aggregators such as EAGLE or Europeana without further modification.

Chapter 5

Discussion

5.1 Automation vs. Manual Curation

A core finding of this project is the identification of the functional boundary where general-purpose automation must give way to manual expert intervention. While the `mysql_to_epidoc.py` script successfully generates most of the required structural and metadata framework, certain high-level tasks were reserved for manual post-processing to maintain the versatility and re-usability of the codebase.

5.1.1 In-Text Semantic Date Encoding

One of the primary manual interventions involved the semantic encoding of dates embedded within the inscription text itself. While the script automatically handles metadata dating in the `<teiHeader>` using columns like `date_display_en`, identifying dates within the Armenian diplomatic transcription (e.g., extracting "ՉԻ" as a numerical value) requires a level of contextual awareness that exceeds simple regex patterns.

- **Manual Task:** Researchers manually wrapped specific text segments in `<date>` tags within the `<body>` of the inscription.
- **Rationale:** This ensures that chronological markers are machine-searchable not only as metadata but as internal textual evidence, allowing for sophisticated queries regarding how dates were expressed paleographically over time.

5.1.2 Lemmatization and Linguistic Analysis

Although the script imports the `dhv_to_epidoc` helper to handle basic expansions and abbreviations, full lemmatization of the Armenian text was performed manually in the final XML files.

- **Manual Task:** Adding `lemma` attributes to `<w>` (word) tags to link conjugated forms to their dictionary headwords.
- **Rationale:** Lemmatization is a highly interpretative task in Classical Armenian studies. By keeping this out of the general script, the code remains adaptable to other Armenian

sub-corpora without imposing a specific, and potentially controversial, linguistic model.

5.1.3 Punctuation and Paleographic Refinement

Certain paleographic nuances, such as the transition from double dots to triple dots as punctuation markers, were handled in the final curation phase.

- **Manual Task:** Refining punctuation marks to reflect specific regional or chronological scribal habits that were simplified during initial database entry.
- **Rationale:** The database was designed for the rapid ingestion of raw tabular data. High-fidelity paleographic representation often requires a visual check against the physical artifact or high-resolution photograph, a step logically positioned after the initial automated export.

5.2 Scalability and Workflow Reproducibility

The decision to maintain a ”general-purpose” script despite these manual requirements was strategic. By not ”over-fitting” the script to the specific requirements of the ArtsakhEpiC dataset, the tools developed here remain immediately applicable to other epigraphic collections.

The following table summarizes the distribution of labor between the automated pipeline and the scholarly editor:

Automated Tasks (Script)	**Manual Tasks (Researcher)**
Metadata mapping and URN linking	Lemmatization of words
Authority file generation	Semantic date tagging within text
Structural TEI/EpiDoc layout	Paleographic punctuation refinement
Bibliography association	Fine-tuning of fragmented text layouts

Table 5.1
Comparison of Automated and Manual Transformation Tasks

Ultimately, this hybrid workflow ensures that the *AutoEpiDoc* project achieves both the breadth required for a large-scale corpus and the depth required for rigorous research.

Chapter 6

Conclusions

The *AutoEpiDoc* project has successfully designed and implemented a scalable, semi-automated pipeline for the digital preservation and encoding of Armenian epigraphic heritage. By transitioning from a fragmented collection of researcher-curated spreadsheets to a standardized, interoperable EpiDoc-XML corpus, this work provides a robust technical foundation for the *ArtsakhEpiC* and *ArmEpiC* initiatives.

6.1 Summary of Achievements

The development of this pipeline addressed several critical challenges in the digitization of cultural heritage:

- **Data Normalization:** Through the use of `csv_to_mysql.py`, the project established a ”Single Source of Truth” within a relational database, ensuring that thousands of records are managed with strict type safety and cross-referential integrity.
- **Semantic Standardization:** The generation of unified authority lists for materials, places, and monuments ensures that the Armenian corpus adheres to international Linked Open Data (LOD) standards, including alignment with Getty AAT and EAGLE vocabularies.
- **Technical Fidelity:** The encoding process successfully managed complex domain-specific requirements, such as bilingual metadata (Armenian and English) and the preservation of the Armenian Era calendar alongside Gregorian dates.

6.2 Impact and Future Work

The hybrid workflow presented in this report, combining automated structural generation with expert philological curation, represents a balanced approach to the digitization of ancient texts. It allows researchers to bypass the repetitive labor of XML tagging, focusing instead on the higher-level scholarly tasks of lemmatization and paleographic analysis.

Moving forward, the modularity of the Python scripts ensures that this pipeline can be adapted for other regional sub-corpora of the Armenian Epigraphic Corpus. Future iterations of the system could integrate:

1. **Full automation:** Merging of the scripts into a general software with graphical user interface to ensure an even easier usage.
2. **Advanced NLP:** The implementation of automated lemmatization tools for Classical Armenian (Grabar) to reduce the manual curation threshold currently required for linguistic analysis.
3. **Web Publication:** The direct ingestion of the generated EpiDoc files into an online portal (e.g., using EFES - EpiDoc Front-End Services) to provide public access to the digitized heritage of Artsakh.

In conclusion, *AutoEpiDoc* demonstrates that a modular, script-based approach to epigraphy not only increases the efficiency of digital corpus creation but also enhances the scholarly rigor and long-term sustainability of the records themselves. By encoding these stones into a machine-readable format, we ensure that the historical and cultural voices of Armenian inscriptions remain accessible and preserved for the digital age.

Bibliography

[1] EpiDoc: Structure of an EpiDoc edition. (2025). *Version 9.7*. Available at: `https://epidoc.stoa.org/gl/latest/supp-structure.html`.

[2] Bayer, M. (2025). *SQLAlchemy: The Python SQL Toolkit and Object Relational Mapper*. Available at: `https://www.sqlalchemy.org/`.

[3] The Pandas Development Team. (2025). *pandas-dev/pandas: Pandas 2.2.0*. Zenodo. Available at: `https://pandas.pydata.org/`.

[4] Lundh, F. (2025). *The ElementTree XML API*. Python Software Foundation. Available at: `https://docs.python.org/3/library/xml.etree.elementtree.html`.

[5] Python Software Foundation. (2025). *xml.dom.minidom — Lightweight DOM implementation*. Available at: `https://docs.python.org/3/library/xml.dom.minidom.html`.

Appendix A : ART0001.xml

```
<?xml-model href="https://www.stoa.org/epidoc/schema/9.7/tei-epidoc.rng"
            type="application/xml"
            schematypens="http://relaxng.org/ns/structure/1.0"?>
        <?xml-model href="https://www.stoa.org/epidoc/schema/9.7/tei-epidoc
    .rng"
            type="application/xml"
            schematypens="http://purl.oclc.org/dsdl/schematron"?>
<TEI xmlns:space="preserve" xmlns="http://www.tei-c.org/ns/1.0" xml:lang="
    en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title xml:lang="eng">Placeholder for title of document in english
    (provided later)</title>
        <title xml:lang="hy">Արձանագրության վերնագիրը հայերեն</title>
        <respStmt>Placeholder for responsibility statement</respStmt>
      </titleStmt>
      <editionStmt>
        <edition xml:lang="eng" n="1.0">First digital edition (2025-12-12)<
    /edition>
        <edition xml:lang="hy">Առաջին թվային
        հրատարակություն (2025-12-12)</edition>
      </editionStmt>
      <publicationStmt>
        <authority xml:lang="en">ArtsakhEpiC - Regional Corpus of Armenian
    Inscriptions from Artsakh, part of the ArmEpiC (Armenian Epigraphic
    Corpus) hosted by the EPFL Digital Humanities Laboratory (DHLAB)</
    authority>
        <authority xml:lang="hy">Արցախէպիկ (ArtsakhEpiC) հայկական
    արձանագրությունների տարածաշրջանային
    հավաքածու՝ընդգրկված Հայէպիկ (ArmEpiC) համահայկական թվային կորպուսի մեջ՝
    տեղակայված՝ Լոզանի Ֆեդերալ Պոլիտեխնիկական Ինստիտուտի Թվային Հումանիտար
    Գիտությունների Լաբորատորիայում (DHLAB)</authority>
        <idno type="filename">ART0001.xml</idno>
        <idno type="armepic">urn:armepic:artsakh:ins:ART0001</idno>
        <availability>
          <license target="https://creativecommons.org/licenses/by-nc/4.0/"
    >CC BY-NC 4.0</license>
          <p xml:lang="en">This record may be freely reused for non-
    commercial research and teaching purposes with proper attribution.</p>
          <p xml:lang="hy">Այս գրառումը կարելի է ազատորեն օգտագործել ոչ
    առևտրային հետազոտական և կրթական նպատակներով՝ պատշաճ հղմամբ:</p>
        </availability>
      </publicationStmt>
```

```
31      <sourceDesc>
32        <collection xml:id="urn:armepic:coll:artsakhepic" xml:lang="en">
    ArtsakhEpiC - Corpus of Armenian Inscriptions from Artsakh</collection>
33        <collection xml:id="urn:armepic:coll:armepic" xml:lang="en">ArmEpiC
     - Armenian Epigraphic Corpus</collection>
34        <msDesc xml:id="ms_Gandzasar_Monastic_complex__monastery">
35          <msIdentifier>
36            <repository ref="urn:armepic:mon:MON0002">
37              <objectName xml:lang="en">Gandzasar Monastic complex /
    monastery</objectName>
38              <objectName xml:lang="hy">Գանձասար վանական համալիր 
    վանք</objectName>
39            </repository>
40          </msIdentifier>
41          <msPart xml:id="ms_Gavit_Narthex">
42            <msIdentifier>
43              <repository ref="urn:armepic:mon:MONPART0010">
44                <objectName xml:lang="en">Gavit (Narthex)</objectName>
45                <objectName xml:lang="hy">Գավիթ</objectName>
46              </repository>
47            </msIdentifier>
48          </msPart>
49          <physDesc>
50            <objectDesc>
51              <supportDesc>
52                <support>
53                  <objectType ref="urn:armepic:objecttype:OBJ0006" xml:lang
    ="en">lintel</objectType>
54                  <objectType xml:lang="hy">բարավոր</objectType>
55                  <material ref="urn:armepic:material:MAT0001" xml:lang="en
    ">tuff</material>
56                  <material xml:lang="hy">տուֆ</material>
57                  <rs type="technique" ref="urn:armepic:technique:TEC001"
    xml:lang="en">carved</rs>
58                  <rs type="technique" xml:lang="hy">փորագիր</rs>
59                </support>
60                <condition/>
61              </supportDesc>
62              <layoutDesc>
63                <layout xml:lang="hy">Հյուսիսային մուտքի կիսակամար քարին
    արտաքուստ</layout>
64                <layout xml:lang="eng">On the stone semi-arch of the
    northern entrance, from the outside</layout>
65              </layoutDesc>
66            </objectDesc>
67            <handDesc>
68              <handDesc>
69                <handNote xml:id="hand_ART0001" scriptRef="
    urn:armepic:script:SCR002">
70                  <term xml:lang="en">Bolorgits Erkatagir</term>
71                  <term xml:lang="hy">Բոլորգիծ երկաթագիր</term>
72                </handNote>
73              </handDesc>
74            </handDesc>
75          </physDesc>
76          <history>
```

```
77            <origin>
78              <origPlace>
79                <placeName ref="urn:armepic:place:PLC0009">
80                  <name xml:lang="eng">Vank</name>
81                  <name xml:lang="hy">Վանք</name>
82                </placeName>
83              </origPlace>
84              <origDate>
85                <date calendar="#cal_gregorian" when="1271">AD 1271</date>
86              </origDate>
87            </origin>
88            <provenance type="found">
89              <placeName ref="urn:armepic:place:MON0002">
90                <name xml:lang="en">Gandzasar Monastic complex / monastery<
    /name>
91                <name xml:lang="hy">Գանձասար վանական համալիր  վանք</name>
92              </placeName>
93            </provenance>
94            <provenance type="observed">
95              <placeName ref="urn:armepic:place:MON0002">
96                <name xml:lang="en">Gandzasar Monastic complex / monastery<
    /name>
97                <name xml:lang="hy">Գանձասար վանական համալիր  վանք</name>
98              </placeName>
99            </provenance>
100         </history>
101       </msDesc>
102     </sourceDesc>
103    </fileDesc>
104    <profileDesc>
105      <textClass>
106        <keywords>
107          <term xml:lang="en" ref="urn:armepic:inscriptiontype:INSCT0001">
    donation / dedication</term>
108          <term xml:lang="hy">նվիրատվական</term>
109        </keywords>
110      </textClass>
111    </profileDesc>
112  </teiHeader>
113  <facsimile>
114    <surface xml:id="surf_ART0001"/>
115  </facsimile>
116  <text>
117    <body>
118      <div type="edition">
119        <ab>
120          <lb n="1"/><g type="decoration" subtype="vertical-dots" quantity=
    "2">:</g>
121          Թիւ
122          <g type="decoration" subtype="vertical-dots" quantity="2">:</g>
123          ՉԻ
124          <num value="1271">1271</num>
125          <lb n="2"/>Կ
126          <hi rend="ligature">ամ</hi>
127          <hi rend="ligature">աւ</hi>
128          <hi rend="ligature" xml:id="lig1">ն</hi>
```

```
          <expan>
            <abbr>
              <hi rend="ligature" xml:id="lig2">Ա</hi>
            </abbr>
            <ex>ստուծո</ex>
            <abbr>յ</abbr>
          </expan>
          <join xml:id="j1" result="ligature" target="#lig1 #lig2"/>
          , ես՝ Յոհանէս,
          <lb n="3"/><hi rend="ligature">որ</hi>
          դի Իւանէի՝ առաջն
          <hi rend="ligature">որ</hi>
          դ
          <expan>
            <abbr>Ս</abbr>
            <ex>ուր</ex>
            <abbr>բ</abbr>
          </expan>
          ու
          <lb n="4"/>խտիս
          <hi rend="ligature">Գան</hi>
          ձաս
          <hi rend="ligature">ար</hi>
          ա հր
          <hi rend="ligature">ամ</hi>
          <hi rend="ligature">ան</hi>
          <hi rend="ligature">աւ</hi>

          <expan>
            <abbr>տ</abbr>
            <ex>եառ</ex>
            <abbr>ն</abbr>
          </expan>
          Խաչ
          <lb n="5"/>ինո Աթ
          <hi rend="ligature">աբ</hi>
          <hi rend="ligature">եկի</hi>
          ն ի յիմ հալալ
          <hi rend="ligature">ար</hi>
          դե
          <hi rend="ligature">ան</hi>
          ց գ
          <hi rend="ligature">նե</hi>
          ց
          <lb n="6"/>ի զվ
          <hi rend="ligature">ար</hi>
          <hi rend="ligature">դան</hi>
          աթաղս,
          <hi rend="ligature">մին</hi>
          չ
          <hi rend="ligature">որ</hi>
          եց եզ
          <hi rend="ligature">ն ե</hi>
          ւ այլ ընծ
          <lb n="7"/>էք տվի ի
          <expan>
```

```
            <abbr>Ս</abbr>
            <ex>ուր</ex>
            <abbr>Ք</abbr>
          </expan>
          Կաթողիկէս□
          <hi rend="ligature">մի</hi>
          <hi rend="ligature">աբ</hi>
          <hi rend="ligature">ան</hi>
          քս տվին զՈհ
          <hi rend="ligature">ան</hi>
          ու□ զ
          <lb n="8"/><hi rend="ligature">Ակ</hi>
          <hi rend="ligature">որ</hi>
          ա տ
          <hi rend="ligature">աւ</hi>
          նն զ
          <hi rend="ligature">ամ</hi>
          էն
          <hi rend="ligature">եկե</hi>
          <hi rend="ligature">ղեց</hi>
          իք
          <hi rend="ligature">ս ի</hi>
          նձ
          <hi rend="ligature">պա</hi>
          տ
          <hi rend="ligature">ար</hi>
          <hi rend="ligature">ազ</hi>
          . ով խ
          <lb n="9"/><hi rend="ligature">ափան</hi>
          է□
          <hi rend="ligature">դա</hi>
          տի յ
          <expan>
            <abbr>Ա</abbr>
            <ex>ստուծո</ex>
            <abbr>յ</abbr>
          </expan>
          <g type="fullstopr">:</g>
        </ab>
      </div>
      <div type="translation">
        <p xml:lang="hy">Հայոց □□□ □□□□□□ թվին□ Աստծո կամքով□ ես□ Հովհաննես□ որդի Իվանեի՝ առաջնորդ Գանձասարի Սուրբ ուխտի□ Խաչենի տեր Աթաբեկի հրամանով□ իմ արդար միջոցներով գնեցի Վարդանաթաղը□ չորս եզ և այլ ընծաներ տվեցի Սուրբ Կաթողիկեիս: Միաբաններն □ինձ□ բոլոր եկեղեցիներում պատարագ տվեցին Հովհաննի և Հակոբի տոնին□ Ով խափանի□թող դատվի Աստծուց:</p>
      </div>
      <div type="translation">
        <p xml:lang="eng">In the year 720 (1271) of the Armenian calendar, by the will of God, I, Hovhannes, son of Ivane, by the order of the leader of the Holy Order of Gandzasar, the Atabek of Khachen, bought Vardanatagh with my own means, gave four oxen and other offerings to the Holy Catholicos. The monks (for me) celebrated the liturgy in all the churches on the feast of John and Jacob. Whoever interferes, let him be judged by God.</p>
      </div>
```

```
231       <div type="commentary">
232         <p xml:lang="hy">Փորագրությունը համաչափ է՝ արված վարպետ գրչի
    ձեռքով:</p>
233         <p xml:lang="eng">The carving is symmetrical and executed by a
    master scribe.</p>
234       </div>
235       <div type="bibliography">
236         <reference n="1" source="urn:armepic:bibl:BIBL_002" ana="topograph"
    />
237         <reference n="2" source="urn:armepic:bibl:BIBL_003" ana="topograph"
    />
238         <reference n="3" source="urn:armepic:bibl:BIBL_004" ana="topograph
    (reprint)"/>
239         <reference n="4" source="urn:armepic:bibl:BIBL_006" ana="monograph"
    />
240         <reference n="5" source="urn:armepic:bibl:BIBL_005" ana="monograph"
    />
241         <reference n="6" source="urn:armepic:bibl:BIBL_001" ana="corpus"/>
242       </div>
243     </body>
244   </text>
245 </TEI>
```