From Online Resource to Reusable Dataset: The Dataset of Byzantine Book Epigrams

Paulien Lemay[a*], Klaas Bentein[a], Frederic Lamsens[b], Joren Six[b], Kristoffel Demoen[a], Els Lefever[c]

[a]Department of Linguistics, Ghent University, Ghent, Belgium

[b]Ghent Centre for Digital Humanities, Ghent University, Ghent, Belgium

[c]Language and Translation Technology (LT3), Ghent University, Ghent, Belgium

[*]Corresponding author: Paulien Lemay; paulien.lemay@ugent.be

Data paper

Abstract

This paper presents a structured, machine-readable dataset of approximately 13,000 Byzantine book epigrams, mostly dating from the 11th to 15th centuries and written in Byzantine Greek. The dataset is derived from the Database of Byzantine Book Epigrams, an ongoing Ghent University project that provides a web interface for exploring the epigrams but does not support large-scale exports or cross-referencing of metadata. Our pipeline extracts and normalizes information from Elasticsearch and PostgreSQL, stores it in SQLite, and exports the results to Zenodo. By structuring the data, it becomes easier to use in computational and AI-driven research and allows intuitive linking of metadata with information from other sources.

Keywords: Byzantine manuscripts; book epigrams; digital humanities; Ancient Greek; historical linguistics

Author roles: Paulien Lemay: investigation, software, writing – original draft, writing – review & editing; Klaas Bentein: writing – review & editing; Frederic Lamsens: conceptualization, software; Joren Six: conceptualization, software, writing –review & editing; Kristoffel Demoen: writing – review & editing; Els Lefever: writing – review & editing;

1 Overview

Byzantine manuscripts frequently contain short metrical texts known as book epigrams. These writings, while integrated into the manuscript, are usually visually distinguished from the main text, for example by being placed in the margins or between sections. They document aspects of the manuscript's production, ownership, and use, and may identify scribes or patrons, comment on the text's content, or address readers directly. Book epigrams exhibit considerable variation in length and literary sophistication, ranging from brief annotations to carefully composed poems, and function as a form of paratext, mediating the relationship between the main text and its material and social context (Bernard & Demoen, 2019).

The Database of Byzantine Book Epigrams, hosted at Ghent University, is an online platform created to systematically collect, study, and publish these poems. Each epigram is represented in the database as an *Occurrence*, with related textual variants organized as *Types*, allowing users to explore both individual texts and their relationships. While the initial corpus was based on published catalogues, ongoing work supplements it with new transcriptions and critical editions derived from direct manuscript consultation (Ricceri et al., 2023).

Although the database has been valuable for philological and linguistic research, the underlying data was not previously available in a format suitable for computational analysis. While an earlier Zenodo release (Demoen et al., 2023) provides a static SQL dump of the PostgreSQL database, it is neither fully updated nor properly structured: foreign key relationships are incomplete, documentation is limited, and the dump cannot be used directly with the current application. The dataset presented here addresses these limitations by providing a fully structured, machine-readable version of the database, enabling reproducible research, large-scale analyses, and interoperability with other digital resources.

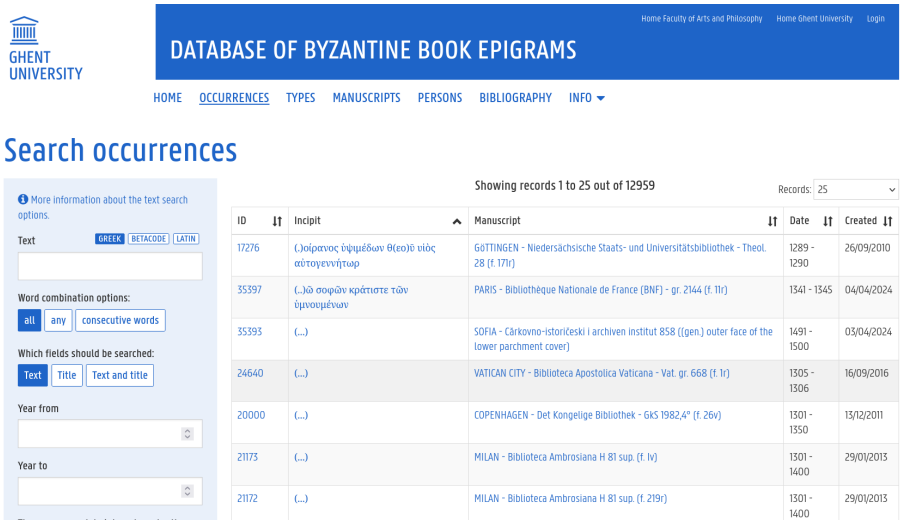

Figure 1: The Database of Byzantine Book Epigrams as available on dbbe.ugent.be

Repository location DOI: 10.5281/zenodo.19435533

Context The dataset was refactored from the original database as part of a doctoral research project investigating the application of deep learning methods to historical Greek texts, conducted within the framework of the ANNOPHIS project[1].

2 Method

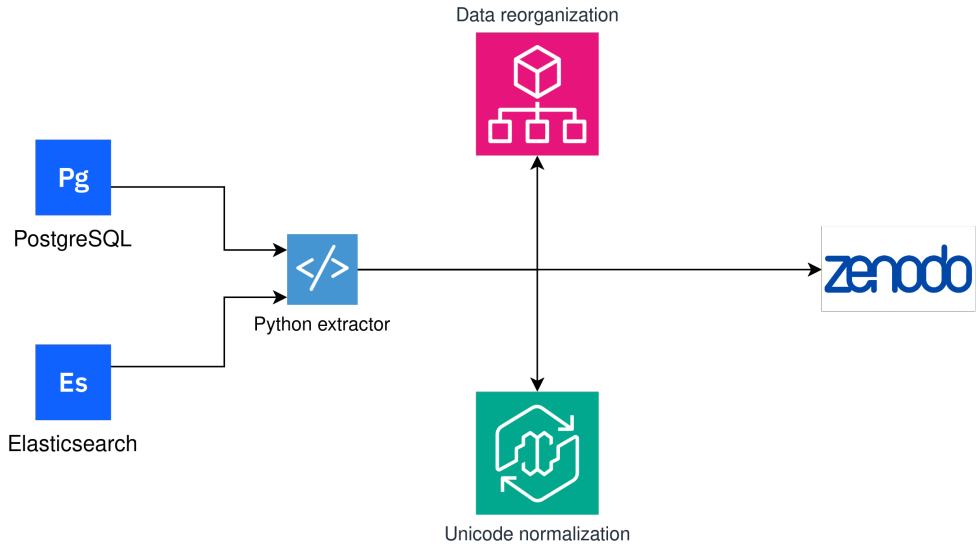

Figure 2: Overview of the automated workflow that runs weekly to archive data from the Database of Byzantine Book Epigrams

Steps The legacy PostgreSQL dataset includes several schemas: `public`, `migration`, `logic`, `data`, and some additional schemas related to specific research projects. The `public` and `migration` schemas contain outdated and unused data, while the `logic` schema stores website-specific information. For the new SQLite database, we focus solely on the core `data` schema to maintain a clean and manageable structure.

The live application employs a custom Elasticsearch setup that indexes data from the data schema, thereby shielding the core database from heavy load. Accordingly, we first extract all data available in Elasticsearch and subsequently retrieve any remaining information directly from PostgreSQL to ensure a complete migration.

[1] https://www.annophis.ugent.be/

During this process, we also restructure the data to create a more logical and coherent schema, fixing foreign keys and organizing tables so that relationships are explicit and consistent. After gathering and restructuring the data, textual content is normalized to Unicode NFC[2] to ensure consistent representation of Greek characters. Scribal ligatures and abbreviations are resolved during transcription according to editorial conventions adopted by the DBBE team and are not subject to a separate processing pipeline[3]. The final dataset is stored in SQLite, a lightweight, self-contained format suitable for long-term archiving, and is then released on Zenodo, which ensures reproducibility and public access.

A public/private flag is used to control data visibility in two ways. On one hand, it allows private comments (e.g., to-do notes or review remarks) attached to Occurrences, Types, Manuscripts, and Persons to be excluded from specific outputs, such as the Zenodo publication. On the other, it allows entire entities to be marked as private when they are not suitable for release (e.g., work in progress, incomplete records, or material outside the project's chronological scope). Public exports include only non-private data, while the full structured dataset, including private entries, is preserved separately for internal use and disaster recovery.

The workflow is orchestrated by a weekly scheduled, containerized task that automates the extraction, processing, and upload of the SQLite dataset to Zenodo. If no changes are detected, no new version is released. Content-level changes (additions, updates, deletions) result in a new release and can be identified by comparing successive Zenodo versions. Schema-level changes in either the PostgreSQL database or the SQLite structure invalidate the transformation logic, triggering failure notifications via a health check service and requiring manual updates to the pipeline and documentation.

Reproducibility is ensured by linking each Zenodo release to the corresponding timestamped state of both the version-controlled pipeline code and the PostgreSQL backup used for extraction. These backups follow a rolling retention schedule (daily for one week, weekly for one month, monthly for one year) and can be accessed upon request.

Quality Control To verify the migration's completeness, we implement extensive tests based on the original PostgreSQL data schema. Each source table is checked using a dedicated Python script to ensure all rows and columns are accurately represented in the new SQLite database. A very limited number of tables and columns are intentionally excluded from the migration; all such exceptions are clearly documented and justified in the project README. All scripts to run the migration can be accessed on `https://github.com/GhentCDH/DBBE-archiver`.

3 Dataset Description

Occurrences	**Types**	**Manuscripts**	**Persons**	**Bibliographical entities** (articles, books, blogs, ...)
Metre Genre Keywords Acknowledgements Text Status Bibliography Persons	Metre Genre Keywords Acknowledgements Translations Bibliography Persons	Location Content Library Persons	Occupation Self designation Acknowledgements	

Table 1: High-level overview of the new dataset structure. Column headers denote main conceptual tables; listed items represent related subordinate tables. The overview captures the core structure without detailing all components.

Repository name Zenodo, GitHub

Object name Dataset of Byzantine Book Epigrams

Format names and versions SQLite

Creation dates The first phase of the Database of Byzantine Book Epigrams began in 2010, and the database continues to be updated with newly added epigrams to the present day.

[2]NFC (Normalization Form Composed) is a Unicode normalization form in which a base character plus one or more diacritics (e.g., e + ´) is combined into a single precomposed character (e.g., é), ensuring visually identical text is represented consistently at the binary level

[3]Ligatures are expanded into separate characters. Abbreviations are written in full in Types, while in Occurrences they are expanded and indicated in parentheses.

Dataset creators Kristoffel Demoen (supervisor), Gilbert Bentein (collaborator), Klaas Bentein (collaborator), Floris Bernard (collaborator), Julián Bértola (collaborator), Julie Boeten (collaborator), Mathijs Clement (collaborator), Cristina Cocola (collaborator), Eline Daveloose (collaborator), Sien De Groot (collaborator), Pieterjan De Potter (developer), Ilse De Vos (collaborator), Maxime Deforche (collaborator), Evelyne Diels (collaborator), Kyriaki Giannikou (collaborator), Quinten Goethals (collaborator), Anna Gregoriani (collaborator), Juan Bautista Juan López (collaborator), Krystina Kubina (collaborator), Frederic Lamsens (collaborator), Eleonora Lauro (collaborator), Hanne Lauwers (collaborator), Paulien Lemay (collaborator), Renaat Meesters (collaborator), Marjolein Morbé (collaborator), Delphine Nachtergaele (collaborator), Marthe Nemegeer (collaborator), Joachim Nielandt (collaborator), Mace Ojala (collaborator), Lisa-Lou Péchillon (collaborator), Raf Praet (collaborator), Rachele Ricceri (collaborator), Anne-Sophie Rouckhout (collaborator), Francesca Samorì (collaborator), Jeroen Schepens (collaborator), Ricarda Schier (collaborator), Febe Schollaert (collaborator), Lev Shadrin (collaborator), Nina Sietis (collaborator), Joren Six (collaborator), Dimitrios Skrekas (collaborator), Colin Swaelens (collaborator), Maria Tomadaki (collaborator), Piril Us MacLennan (collaborator), Sarah-Helena Van den Brande (collaborator), Merel Van Nieuwerburgh (collaborator), Lotte Van Olmen (collaborator), Noor Vanhoe (collaborator), Nina Vanhoutte (collaborator), Grigory Vorobyev (collaborator).

Affiliation of all contributors (at the time of data development): Ghent University, Ghent, Belgium.

Language Greek; English for metadata

License CC-BY (4.0)

Publication date 2026-04-06.

4 Reuse Potential

Project-internal experiments with language models show promising results on the database for tasks such as part-of-speech tagging and lemmatization (Swaelens, 2025). By providing the complete corpus in a structured, machine-readable SQLite format with normalized Greek texts, the dataset now makes it easier for researchers with diverse expertise to explore a wider range of LLM applications. These include classifying epigrams by type, genre, or metre, performing linguistic analyses, and extracting named entities while leveraging the fully linked metadata. Initial tests also indicate that large language models can be used to query the data in plain English by pointing to the path of the SQLite file, opening up new possibilities for interactive and flexible exploration of the corpus. This release removes previous barriers of limited web access or ad hoc data requests, facilitating both generative modeling and structured computational analyses.

Structuring the database also enables the generation of comprehensive statistics and a wide range of flexible data representations. Firstly, constructing graph-based representations becomes substantially easier with the explicit organization of entities and relations. A project collaborator has previously demonstrated the potential of this approach to explore orthographical similarity between occurrences, types, and word variants (Deforche et al., 2024), using the current database. With the new structured data, similar graph-based analyses can now be conducted more efficiently and are readily accessible to researchers. Secondly, the dataset also supports diverse analytical and visual approaches, allowing researchers to select the representations best suited to address specific research questions. For example, it is now possible to examine the distribution of epigrams across manuscripts, collections, or chronological periods (Figure 3), a task that previously required ad hoc requests to the project team due to the complexity of the underlying data.

Providing clear identifiers for manuscripts and persons enables the dataset to be enriched with information from external resources such as Pinakes[4] or Biblissima[5]. This enrichment facilitates cross-corpus analyses and improves interoperability with other datasets.

Despite these advantages, certain limitations should be acknowledged. The database remains a growing corpus. Therefore, quantitative analyses or coverage-based conclusions should be interpreted with care. Additionally, some widely used external identifiers, such as those from Pleiades[6], are not yet included, which may constrain interoperability with other linked-data projects. Nevertheless, by providing structured, machine-readable data and harmonized identifiers, the dataset significantly lowers barriers for reuse, enabling aggregation, validation, teaching, and collaborative research across multiple scholarly domains.

[4] https://pinakes.irht.cnrs.fr/
[5] https://projet.biblissima.fr/
[6] https://pleiades.stoa.org/

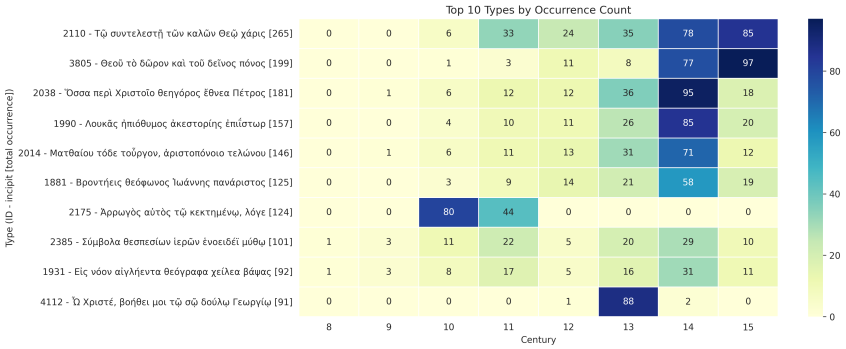

Figure 3: Demonstration of the types of summary statistics that can be produced from the structured dataset.

Acknowledgements

This project is supported by GhentCDH (the Ghent Centre for Digital Humanities), a core facility at Ghent University specialised in providing consultancy on Digital Humanities projects.

Funding Statement

This work was supported by the Research Foundation Flanders (FWO) under grant I005524N and by the Special Research Fund (BOF) under grant GOA.2021.0006.03.

Competing interests

The authors have no competing interests to declare.

References

Bernard, F., & Demoen, K. (2019). Byzantine book epigrams. In *A companion to Byzantine poetry* (pp. 404–429). Brill.

Deforche, M., De Vos, I., Bronselaer, A., & De Tré, G. (2024). A hierarchical orthographic similarity measure for interconnected texts represented by graphs. *Applied Sciences*, *14*(4). Retrieved from https://www.mdpi.com/2076-3417/14/4/1529 DOI: 10.3390/app14041529

Demoen, K., Bentein, G., Bentein, K., Bernard, F., Bértola, J., Boeten, J., ... Vanhoutte, N. (2023, February). *Database of byzantine book epigrams.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7682523 DOI: 10.5281/zenodo.7682523

Ricceri, R., Bentein, K., Bernard, F., Bronselaer, A., De Paermentier, E., De Potter, P., ... others (2023). The database of byzantine book epigrams project: Principles, challenges, opportunities. *Journal of Data Mining & Digital Humanities*(Project presentations). Retrieved from https://jdmdh.episciences.org/11516 DOI: 10.46298/jdmdh.10244

Swaelens, C. (2025). *Verse by verse: Modelling semantic similarity in Byzantine Greek poetry* (Unpublished doctoral dissertation). Ghent University.