Building a Corpus and Database for Rare and Undeciphered Scripts

Beáta Megyesi[1], **Rune Rattenborg**[2], **Benedek Láng**[3], **Michelle Waldispühl**[4], **and Mihály Héder**[5]
[1]Stockholm University, Sweden
[2]Lund University, Sweden
[3]Eötvös Loránd University (ELTE), Hungary
[4]University of Oslo, Norway
[5]Budapest University of Technology and Economics / HUN-REN SZTAKI, Hungary
{ beata.megyesi@ling.su.se, rune.rattenborg@klass.lu.se, lang.benedek@gtk.elte.hu, michelle.waldispuhl@ilos.uio.no, mihaly.heder@sztaki.hu }

Abstract

Historical sources written in rare or undeciphered scripts represent an immense but underexploited part of the world's cultural and linguistic heritage. Their study is often hindered by fragmentary preservation, non-standard symbol systems, and the absence of interoperable digital resources. While recent advances in imaging, transcription, and computational analysis have improved access to historical texts, most tools rely on large quantities of labeled data and standardized encodings, requirements that are rarely met for rare or unknown writing systems. This paper presents the design and methodology of a new corpus and database dedicated to rare and undeciphered scripts worldwide. The resource integrates high-quality images, transliterations, transcriptions, linguistic annotations, and metadata within a unified data model tailored for low-resource and non-standard scripts. By adhering to FAIR principles and existing standards for linguistic and cultural heritage data, the database enables reproducible, interdisciplinary research across philology, linguistics, cryptology, and computer science. The paper outlines the data collection and digitization workflow, describes the metadata and database architecture, and demonstrates applications in analysis and decipherment.

Keywords: decipherment, historical writing systems, rare scripts

1. Introduction

Historical sources written in rare or undeciphered scripts are among the most valuable yet least accessible parts of the world's cultural record. They preserve unique evidence of linguistic diversity, intellectual history, and cultural exchange, but remain systematically underused because they are difficult to access, standardize, and analyze. Examples of such writings include some of the most famous enigmas in the history of writing: the Indus Valley inscriptions from South Asia, the Rongorongo tablets of Easter Island, the Vinča signs of prehistoric southeastern Europe, the Linear A tablets from the Eastern Mediterranean and Proto-Elamite tablets from the ancient Near East, and the enigmatic Voynich manuscript from early modern Europe. Although each belongs to a distinct script tradition, material context, and historical setting, all present similar challenges: fragmentary evidence, uncertain linguistic affiliation, and a lack of standardized encoding or digital representation.

Traditional philological and epigraphic methods rely on expert knowledge, comparative materials, and well-established reference corpora. While these approaches have produced major breakthroughs in well-documented traditions, they encounter severe limitations when evidence is fragmentary, symbol inventories are idiosyncratic, or the underlying language and script remain unknown.

Digital technologies have substantially improved the documentation and legibility of historical materials. High-resolution imaging, multispectral enhancement, and handwritten text recognition (HTR) now form part of the standard toolkit for manuscript and epigraphic studies. Yet these tools depend on abundant labeled data and stable script encodings, conditions rarely met for rare or non-standard writing systems. In many cases, even a few hundred annotated pages are unattainable, and even basic Unicode (The Unicode Consortium, 2019) coverage may be lacking. Consequently, many historically and linguistically important sources remain outside the reach of computational methods and large-scale linguistic analysis.

Recent advances in historical cryptology and computational decipherment illustrate both the potential and the challenges of algorithmic assistance. Techniques such as language modeling, frequency analysis, and heuristic search can accelerate interpretation, but existing solutions remain specialized, fragmented, or dependent on expert intervention. There is still no generic, end-to-end workflow capable of supporting the full range of tasks required to process rare or undeciphered scripts—from image segmentation and symbol inventory creation to linguistic interpretation and contextualization.

A systematic, interoperable collection and database of historical sources with rare and un-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 184–196.

known writing systems is therefore an essential step forward. Such a resource enables the preservation and comparative study of diverse scripts, supports the development of transferable AI and HTR models under low-data conditions, and provides a foundation for reproducible computational experiments. It also facilitates collaboration across disciplines—linking linguistics, archaeology, history, and computer science—and ensures that cultural heritage materials can be documented and analyzed within FAIR and open-data frameworks.

In this paper, we present the design and methodology of a new corpus and database dedicated to rare and undeciphered scripts. Section 2 reviews related work on the study and digitization of rare and unknown writing systems. Section 3 describes the data collection and digitization workflow, including source types, datasets, image acquisition, transcription, and metadata tailored to rare and undeciphered writings. Section 4 presents the database architecture and its alignment with established standards for linguistic and cultural heritage data. Section 5 discusses collaboration and long-term aims, and Section 6 concludes the paper.

2. Background

The study of historical writings provides an invaluable window into the past, offering insights into the cultures, languages, and intellectual traditions of earlier societies. Traditionally, research on ancient and rare scripts has relied on philological and epigraphic expertise: specialists meticulously examine texts within their linguistic and historical contexts, often focusing on sources from the same region, period, and language. While such methods have produced foundational breakthroughs—from the decipherment of Egyptian hieroglyphs by Champollion (1822) to the work on Linear B by Chadwick (1958) or the Maya script (Coe, 2011)—they remain limited by the fragmentary nature of the data, the scarcity of comparative material, and the time-intensive process of manual analysis.

Over the past few decades, the digital humanities have transformed access to and analysis of historical materials. Digitization initiatives, such as high-resolution imaging, 3D modeling, and optical character recognition (OCR), have made it possible to preserve, share, and annotate ancient texts at scale (Babeu, 2011; Santis and Rossi, 2019; Soriano and Espinosa, 2021). However, these technologies perform best on well-preserved and standardized scripts, and typically fail when confronted with irregular, damaged, or non-standard writings, features that characterize many rare and undeciphered sources. Moreover, variation in metadata standards, annotation practices, and encoding formats limit the interoperability of existing resources, making large-scale comparison and computational analysis difficult.

Current repositories dedicated to artifact corpora vary widely in scope: for instance, the cuneiform corpus contains over 550,000 inscriptions (Streck, 2010; Rattenborg et al., 2023), compared to around 500,000 Latin (Lloris, 2015), 200,000 Greek (Liddel, 2017), 15,000 Ancient South Arabian (Avanzini, 2009), 8,000 Runic (Williams et al., 2022), 6,000 Linear B (Aurora, 2015), and an estimated 20-100,000 texts combined for Egyptian hieroglyphic, hieratic, and demotic inscriptions (Trismegistos, 2025; Thesaurus Linguae Aegyptiae, 2025). Beyond these better-known cases, thousands of sources remain isolated: inscriptions on bone, wood, clay, parchment, and stone, or manuscripts written in unknown symbols or numeral sequences. Many of these represent linguistic traditions yet to be deciphered or even identified.

Efforts to analyze such material have increasingly turned to computational linguistics, cryptology, and AI. Historical cryptology—the systematic study of historical ciphers and secret writings often featuring non-standard symbol systems and low-resource language data (Kahn, 1967; Megyesi et al., 2024a)—has demonstrated how algorithmic methods can complement traditional expertise. The DECRYPT[1] project (Megyesi et al., 2020) pioneered this approach by creating large-scale digital resources for early modern European ciphertexts and cipher keys, including the DECODE cipher database (Megyesi et al., 2019; Héder and Megyesi, 2022) and Megyesi et al. (2024), the HistCorp multilingual historical corpora for European languages (Pettersson and Megyesi, 2018) and (Pettersson and Megyesi, 2024), and advanced cryptanalysis tools such as CrypTool 2 (Esslinger, 2024).

Building on prior research in automated text and cipher analysis, the work presented here broadens the scope from European sources to rare and undeciphered scripts worldwide, integrating AI, computer vision, and linguistic modeling to facilitate transcription, decipherment, and contextual interpretation.

Parallel research in automatic decipherment and transcription has shown the potential of computational approaches. Statistical and machine learning methods have been successfully applied to historical ciphers and ancient scripts (Dou and Knight, 2012; Ravi and Knight, 2011b,a; Knight et al., 2012; Reddy and Knight, 2011, 2012; Hauer and Kondrak, 2016; Lasry, 2018; Yin et al., 2019; Lasry et al., 2020; Leierzopf et al., 2021a,b; Kopal and Waldispühl, 2022; Souibgui et al., 2022; Lasry et al., 2023). Similar methods have been explored for undeciphered scripts such as Luwian hieroglyphs (de Lin and Knight, 2006), Ugaritic (Sny-

[1] de-crypt.org

der et al., 2010), and Iberian (Luo et al., 2021), though large-scale success remains elusive (Ferrara and Tamburini, 2022). Recent breakthroughs, such as the decipherment of Linear Elamite (Desset et al., 2022), Amorite vocabularies (George and Krebernik, 2022), and the Kushan script (Bonmann et al., 2023), highlight the enduring importance of systematic documentation, transcription, and interdisciplinary collaboration.

In parallel, research on script recognition and image enhancement (Hochberg et al., 1997; Easton and Knox, 2016; Fornés et al., 2017; Baró et al., 2019; Chen et al., 2020; Souibgui et al., 2022; Szigeti and Héder, 2022) and neural reconstruction of damaged manuscripts (Fetaya et al., 2020; Assael et al., 2022; Papavassileiou et al., 2023) has advanced rapidly. Platforms such as Transkribus (Tra, 2024) and eScriptorium (eSc, 2025) offer infrastructures for handwritten text recognition, but their reliance on large training datasets makes them unsuitable for rare or low-resource scripts. Even specialized tools such as Fabricius (Fab, 2024), developed for Egyptian hieroglyphs, remain restricted in scope. Consequently, generic and flexible transcription tools that can operate with minimal annotated data are still lacking.

These technological limitations underscore the continued need for human expertise and interdisciplinary cooperation. As Terras (2012, 2016) argues, digital palaeography and crowdsourced manuscript analysis must integrate historical knowledge with computational methods to achieve meaningful progress. The combination of domain expertise from philologists, linguists, archaeologists, computer scientists, and cryptologists offers a powerful model for advancing the study of rare and undeciphered scripts.

The present project builds directly on this interdisciplinary tradition. By creating a global corpus and database that integrates images, transliterations, linguistic annotation, and metadata, and eventually sign inventories, it aims to provide a unified and extensible infrastructure for studying the world's writing diversity. The resource is intended to support both humanistic and computational research, enabling the preservation, comparison, and analysis of rare scripts, and it aligns with the LREC mission to promote FAIR linguistic data. Designed according to the Linguistic Linked Open Data[2] (LLOD) principles and compatible with CLARIN[3] and ELRA[4] infrastructures, it contributes to the long-term goal of sustainable, reproducible, and inclusive language resources.

[2] https://linguistic-lod.org
[3] https://www.clarin.eu
[4] https://www.elra.info

3. The DESCRYPT Corpus

3.1. Scope and Source Types

The DESCRYPT[5] project collects and curates a wide range of materials featuring rare, non-standard, or unknown scripts, including but not limited to (i) ancient inscriptions (e.g., Ajami, Cuneiform, Linear A/B, Old South Arabian, Linear and Proto Elamite, runic corpora), (ii) manuscript traditions with specialized or hybrid notation (e.g., shorthand systems, also called stenographic writing, artificial language schemes, mixed symbol sets), and (iii) encrypted or otherwise encoded documents from later periods (diplomatic, scientific, or private correspondence; political or religious materials). The corpus includes both famous enigmas and lesser-known fragments found on bone, wood, clay, stone, parchment, or paper, whether preserved in archives, libraries, museums, or private collections. Figure 1 illustrates some well-known rare writings.

The first group also encompasses undeciphered and rare historical scripts in natural languages from later epochs (e.g., Rongorongo or the contentious so-called Hungarian runic script), while the second includes early modern artificial language schemes and "common writings" devised by philosophers and linguists such as John Wilkins, Athanasius Kircher, René Descartes, Isaac Newton, Gottfried Wilhelm Leibniz, and Marin Mersenne. These often aimed at constructing universal systems of representation and communication. Alongside such projects, shorthand systems, which virtually disappeared by the twelfth century and were reinvented around 1600, proliferated into numerous schools and national traditions; over four centuries, approximately 450 methods emerged in print, especially in English-speaking regions. Many of these shorthand systems later came to resemble ciphers when their explanatory texts were lost, blurring the distinction between stenography and cryptography. The DESCRYPT corpus therefore brings together a wide spectrum of enigmatic writings and encoded systems, including ciphertexts, cipher keys, and rare alphabets, whether solved, partially deciphered, or unsolved.

3.2. Repositories and datasets

Sources included in DESCRYPT are drawn from open access archaeological, philological, and linguistic research databases, museum and library catalogs, and similar authoritative repositories. As a result of the digital turn in the humanities, cataloging and curation of historical writings in recent decades have built a rich ecosystem of digital re-

[5] descrypt.org

Figure 1: Examples of rare scripts.

sources providing access to millions of records, representing historical and rare writing systems from all over the world (Hockey, 2008; Babeu, 2011). Included datasets are evaluated, documented, and mapped to internal metadata and transcription standards using transparent and fully retraceable procedures, and stored and reused in accordance with relevant licenses and permissions of the source repositories.

Datasets included in DESCRYPT already cover a diverse range of scripts illustrating the broader scope and potential of the resource. Current holdings include ancient inscriptions from Europe and the Middle East, among others cuneiform and related writing systems from Iraq, Syria, Iran and Turkey (Streck, 2010; Rattenborg et al., 2023), Linear B from the Eastern Mediterranean (Aurora, 2015), Rongorongo from the Easter Island, and Runic from Scandinavia (Institutionen för nordiska språk, 2020). Chronologically, records span from some of the earliest known writing systems to the early modern period. The database further contains a variety of shorthand systems and records employing artificial language schemes, predominantly European and dating to the Middle Ages and the early modern era, as well as extensive collections of historical ciphertexts assembled as part of the DECRYPT project within the DECODE database, now being further augmented by the DESCRYPT project.

The sources under study are exemplified in Table 1, which presents examples of ancient and modern historical writings and scripts, while Table 2 illustrates examples of encrypted texts, shorthand systems, and artificial language schemes.

3.3. Acquisition and Imaging

Where feasible, sources are linked to authoritative archival repositories; otherwise, the program undertakes high-resolution digitization following best practices in cultural heritage imaging. The pipeline emphasizes image quality and reproducibility to support downstream processing: dewarping, contrast enhancement, denoising, and illumination normalization are applied cautiously, with original images preserved. When available and relevant, 3D models and multispectral derivatives are associated with each record to aid symbol recovery on challenging surfaces.

Digitization relies on both high-resolution scanning and photography, ensuring that every record can serve as a reliable digital surrogate for scholarly analysis. Existing digital collections prepared by other research communities are not duplicated but are instead referenced and linked to the DESCRYPT infrastructure. Each source is annotated, catalogued, and integrated into the database according to the TEI standard, using a flexible schema capable of representing both ciphers and other rare or unknown writing systems.

The resulting, albeit still growing, corpus thus provides a searchable and standardized set of digital reproductions intended for systematic and comparative study.

3.4. Metadata and Contextualization

Given the global outlook of the DESCRYPT project, the collection and integration of source material requires extensive metadata to support linguistic, historical, and archaeological contextualization, as well as efforts at decipherment and translation. Ro-

Script	Region	Date	Status	Type	Language
Asia					
Cuneiform	Iraq, Syria, Turkey	3200 BCE–100 CE	Deciphered	Logo-syllabic	Sumerian, Akkadian
Proto-Elamite	Iran	3100–2900 BCE	Undeciphered	Logographic	Unknown
Linear Elamite	Iran	2300–1880 BCE	Part. dec.	Logo-syllabic	Elamite
Indus (Harappan)	Pakistan, India	2600–1900 BCE	Undeciphered	Logo-syllabic?	Unknown
Brahmi	India	200-300 BCE	Deciphered	Abugida	Prakrit, Sanskrit
Proto-Sinaitic	Sinai	1850–1550 BCE	Part. dec.	Abjad	Proto-Semitic
Ugaritic	Syria	1400–1200 BCE	Deciphered	Abjad	Ugaritic
Oracle Bone Script	China	1200-700 BCE	Deciphered	Logographic	Archaic Chinese
Khitan (Lg./Sm.)	N. China	920–1190 CE	Part. dec.	Logo-syllabic	Khitan
Tangut	W. China	1036–1227 CE	Largely dec.	Logographic	Tangut
Jurchen	Manchuria	1110–1400 CE	Part. dec.	Mixed	Jurchen
Bactrian/Sogdian var.	C. Asia	100–900 CE	Part. dec.	Alphabetic	Iranian
Africa					
Meroitic	Nubia	300 BCE–400 CE	Part. dec.	Abugida	Meroitic
Libyco-Berber (Tifinagh)	N. Africa	300 BCE–pres.	Part. dec.	Abjad	Berber
Nsibidi	Nigeria/Cam.	pre-1600 CE–pres.	Undeciphered	Proto-writing	Unknown
Europe					
Vinča symb.	Balkans	5300–4000 BCE	Undeciphered	Proto-writing	Unknown
Linear A	Crete	1800–1450 BCE	Undeciphered	Logo-syllabic	Unknown (Minoan)
Linear B	Greece/Crete	1450–1200 BCE	Deciphered	Logo-syllabic	Mycenaean Greek
Cypro-Minoan	Cyprus	1550–1050 BCE	Undeciphered	Syllabic	Unknown
Etruscan	Italy	700–100 BCE	Part. dec.	Alphabetic	Etruscan
Runic (Elder Futhark)	N. Europe	ca. 150–750 CE	Deciphered	Alphabetic	Proto-Norse
Americas					
Olmec (Cascajal)	Mexico	ca. 900 BCE	Undeciphered	Glyphic	Unknown
Zapotec	Oaxaca	500 BCE–500 CE	Part. dec.	Logo-syllabic	Zapotec
Epi-Olmec/Isthmian	S. Mexico	300 BCE–500 CE	Part. dec.	Logo-syllabic	Unknown (Mixe-Zoq.)
Maya (early)	Mesoam.	300 BCE–250 CE	Part. dec.	Logo-syllabic	Mayan
Rongorongo	Polynesia	1800–1900 CE	Undeciphered	Glyphic	Unknown

Table 1: Examples of deciphered, largely deciphered (Largely dec.) partially deciphered (Part. dec.), and undeciphered writing systems across major world regions.

bust metadata protocols are essential for integrating sources drawn from a wide range of writing traditions, documented according to highly diverse scholarly conventions and preserved with varying degrees of digital representation and accessibility.

In addition to automatically generated administrative fields, such as the record identifier, and creation and update timestamps, the metadata model comprises fields for file description, record identification, object description, origin and provenance, correspondence, content, and additional information. The *File description* records the person responsible for the entry, the collector, the record type (e.g., natural script, artificial script, or cipher), and the access mode. *Record identification* includes the name of the record, its current institution, city, country, and source, that is, the geographical and institutional location at which the inscribed object is currently held, for example a library or a museum, together with its source identifier, the URI in the source repository, and license information. *Object description* captures the type and material of the inscribed object, together with its measurements and, where relevant, hand and decoration. *Origin and provenance* fields record the place of origin, including city and region, provenance, provenance type, dating range, and historical or archaeological period. The *Content* field describes the inscription through script, language(s), genre, number of pages, and cipher-related features, while *correspondence* fields allow for the specification of author, sender, and recipient. Bibliography and notes along with acknowledgments are recorded separately as *additional information*.

Wherever possible, the schema normalizes description through controlled vocabularies aligned with established external authorities in linguistics, writing systems, and cultural heritage documentation, thereby strengthening data discovery, compatibility, reuse, and linked open data capability (Bizer et al., 2009; Berners-Lee, 2006; Palma and Megyesi, 2025). Record material, object type, and genre are mapped to the Getty Art and Architecture Thesaurus (Getty Research Institute, 2025); language and script identifiers follow ISO 15924 standards and RFC 5646-based categories and best-practice recommendations associated with the Internet Assigned Numbers Authority (IANA)[6]; and geographical and chronological references may be aligned with resources such as Geonames (Geonames), Pleiades (Pleiades), PeriodO, and Wikidata (Wikimedia Foundation, 2025). At the same time, the model retains free-text subtype and note fields in order to preserve legacy values and to accommodate cases in which controlled vocabularies do not

[6] https://www.iana.org

Category	Region	Date	Nature / Type
Historical ciphertexts			
Voynich manuscript	Central Europe?	15th c. CE	Undeciphered (cipher? language?)
Letters of Mary, Queen of Scots	Scotland/England/France	1586 CE	Homophonic substitution cipher
Papal diplomatic ciphers	Italy (Europe-wide use)	16th–17th c. CE	Homophonic and polyphonic substitution cipher
The Borg cipher	Northern Europe	17th c. CE	Simple substitution cipher
The Copiale cipher	Germany	ca. 1730 CE	Homophonic substitution cipher
Cipher keys / devices			
Alberti cipher disk	Italy	1467 CE	Polyalphabetic disk (rotating)
Cipher keys	Europe	14th–19th c. CE	Digit/symbol mappings for letters/words
Vigenère table	France	16th c. CE	Polyalphabetic key table
Jefferson disk	USA/France	1790s CE	Multi-alphabet cylinder
Ottoman cipher tables	Ottoman Empire	18th–19th c. CE	Bigram/word tables + numerals
Shorthand systems			
Tironian notes	Roman world	1st c. BCE	Latin shorthand symbols
Gabelsberger shorthand	Germany	1817 CE	Cursive phonetic shorthand
Pitman shorthand	United Kingdom	1837 CE	Phonetic, stroke thickness/position
Duployan shorthand	France/Canada	1860s CE	Curvilinear, widely adapted
Gregg shorthand	USA	1888 CE	Cursive, elliptical phonetic system
Japanese sokki (Waseda) shorthand	Japan	1917 CE	Mora-based, indigenous systems
Arabic shorthand (Haddad)	Egypt	1940s CE	Pitman-influenced, RTL
Artificial language schemes			
Marin Mersenne's La science universelle	France	1636 CE	Proposal for a universal science and signs
Francis Lodwick's Common Writing	England	1647 CE	Empirical–lexical language
John Wilkins' philosophical language	England	1668 CE	Taxonomic, a priori lexicon
Cave Beck' Pasigraphie	England	1657 CE	Universal writing, numerical codes for words
George Dalgarno's The Universal Character	Scotland, England	1661 CE	Symbolic writing for communication
G. W. Leibniz's Characteristica universalis	Germany	1666-1700 CE	Universal symbolic language
Solresol	France	1820s CE	Do-re-mi based, universalist

Table 2: Examples of historical ciphertexts, cipher keys/devices, shorthand systems, and artificial language schemes across regions and time periods.

provide sufficient granularity. The schema therefore combines standardization and interoperability with the flexibility required for heterogeneous documentary corpora.

Schemas for rendering records in TEI- and EpiDoc-compatible XML (Consortium, 2023; Elliotte et al., 2006–2022) are tailored to include linked metadata (Gaitanou et al., 2022) on object type, script, material, archaeological and historical context, chronological and geographical indicators of use, relations to other pertinent sources, and decipherment status. Each entry may also include separate components for the associated script and language, facilitating cross-referencing within and beyond the database. The metadata scheme is illustrated in Figure 2. This structured descriptive framework forms the basis for subsequent computational processing and collaborative annotation, while also facilitating structured export and comparative analysis.

4. The DECODE Database

To make the data openly accessible, we created a database that supports search, editing, and the upload of new records.

4.1. Design Principles

The DECODE database is a public, extensible repository that unifies *images*, *transcriptions*, *metadata*, *derived annotations*, and *processing artifacts* (segmentations, symbol clusters, model outputs) within a coherent, standards-based framework.

It evolves the DECRYPT infrastructure (Megyesi et al., 2024a) to accommodate a wider typology of writing systems and integrates language resources and tools for analysis.

The database is built from standard components, including a PHP application backed by a relational database, quasi-standard API interface with the *swagger* framework, mass input modules for UX, the API and from CSV format. The experience derived from previous versions is that it is best to provide several paths for data entry, in small and large quantities, as different users have different working methods and preferences. Four principles guide the design:

1. **Interoperability**: TEI/EpiDoc-compliant metadata structure; persistent identifiers; Geonames Open Data value sets and schema alignment to external catalogs; role-based authorization system built on standard components; and a RESTful API for programmatic access.

2. **Provenance and Traceability**: Records linking raw images, pre-processing steps, potential annotations and model inferences; full audit trails of human and programmatic edits and access; clear licensing and rights; metadata.

3. **Data Generalization**: Support for connecting with AI deployments that can train and evaluate models under low-resource constraints

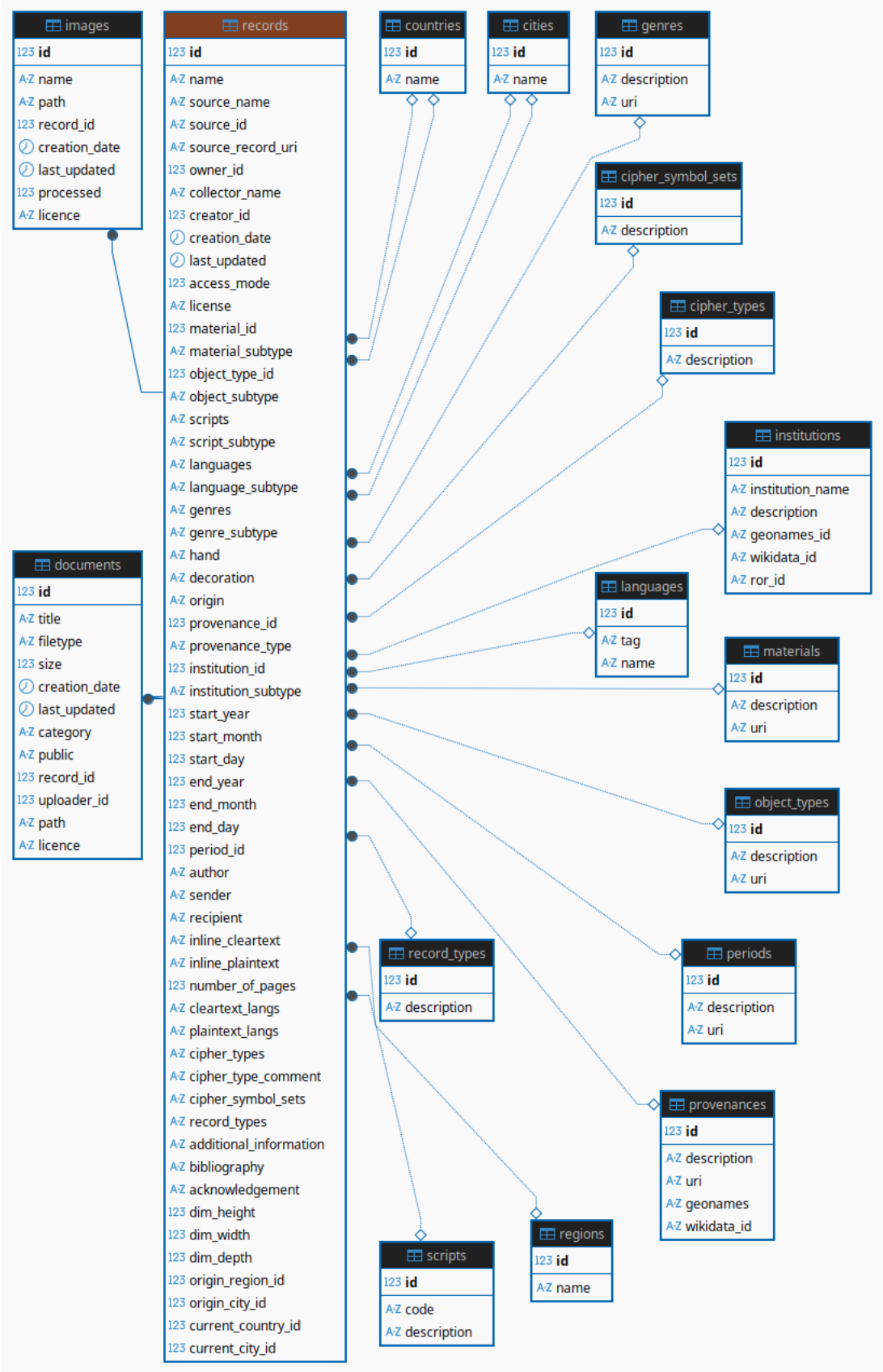

Figure 2: Metadata structure.

(few-shot settings), with benchmark splits for reproducible comparison and shared tasks.

4. **Scholarly Control**: Connectivity to expert-in-the-loop tooling to review, correct, and approve transcriptions and interpretations; mechanisms to represent uncertainty and competing hypotheses. Extensive user testing guide the development and integration of tools, ensuring their effectiveness in real-world applications.

4.2. Content and Schema

Each database record encapsulates:

- **Core Object**: High-resolution images (and, if available, 3D/multispectral derivatives), stable identifiers, source repository links, and rights statements.
- **Descriptive Metadata**: Object type, material, condition; script classification; chronology and provenance; associated persons/places in line with well-known Linked Open Data resources; and bibliographic references.
- **Textual Data**: Diplomatic transcription aligned to images and layout (blocks, lines, regions), with symbol-level segmentation and clustering, plus normalized representations if applicable.
- **Analytical Layers**: Language identification hypotheses; frequency profiles; co-occurrence statistics; cryptanalytic views for encoded sources; alignment to dictionaries or historical language models; and machine-inferred suggestions with confidence scores.
- **Relations and Collections**: Links across items by hand, inscription series, archive, or cipher family; curated sets for teaching, benchmarking, or shared tasks.

Figure 3 shows the current version of the database interface, including several records and selected metadata fields.

5. Collaboration Across Disciplines

The DESCRYPT project is inherently interdisciplinary, bringing together historians, philologists, linguists, cryptologists, computer scientists, and specialists in computer vision and manuscript studies in the joint development of the corpus and database. This combination of expertise ensures that the resource is historically grounded, methodologically robust, and relevant to multiple research communities working with rare writings.

An important objective is for the corpus to function as a *monitor corpus*: a living and continuously expanding resource that can be updated as new material is identified, additional images and descriptions become available, and existing analyses are revised. Because many rare writings survive only in fragmentary, dispersed, or otherwise unique forms, a static corpus would not adequately capture the changing state of documentation and interpretation. The corpus is therefore designed to support continuous enrichment and revision over time.

The database is intended to make this collaborative model operational by allowing domain experts to upload new records, revise existing entries, improve transcriptions and metadata, and contribute alternative readings or interpretations of rare writings and other unique materials. In this way, the platform serves not only as a repository, but also as a shared research environment that supports scholarly exchange and expert-led curation. Such collaboration is essential for possible decipherment of complex historical sources, where progress depends on combining domain-specific knowledge with computational, linguistic, and imaging methodologies (Láng and Megyesi, 2024).

6. Long Term Aims

The long-term goal of DESCRYPT is to develop the corpus and database into a mature research infrastructure for the study of rare, non-standard, and undeciphered writings. Beyond collecting and organizing material, the project aims to support the full research cycle from data acquisition and curation to analysis, interpretation, and reuse. Particular emphasis will be placed on extending coverage, refining data standards for poorly encoded writing systems, and enabling reproducible work across heterogeneous source types.

A central objective is to establish practical transcription and representation conventions for scripts that lack stable editorial or encoding standards. This includes Unicode-compliant solutions where possible, provisional encodings for otherwise unsupported symbols, and explicit links between image-level observations and abstract symbol inventories. Such conventions will make it easier to compare sources across traditions and to create reusable ground truth for computational analysis.

Another long-term aim is to connect the database more directly to analytical workflows. Future development will integrate services for layout analysis, symbol segmentation, transcription support, and exploratory decipherment, allowing researchers to move more efficiently from digitized objects to interpretable textual data. Language resources such as dictionaries, sign inventories, and historical corpora will be linked where relevant, while encoded materials will benefit from cryptanalytic functionality tailored to historical sources.

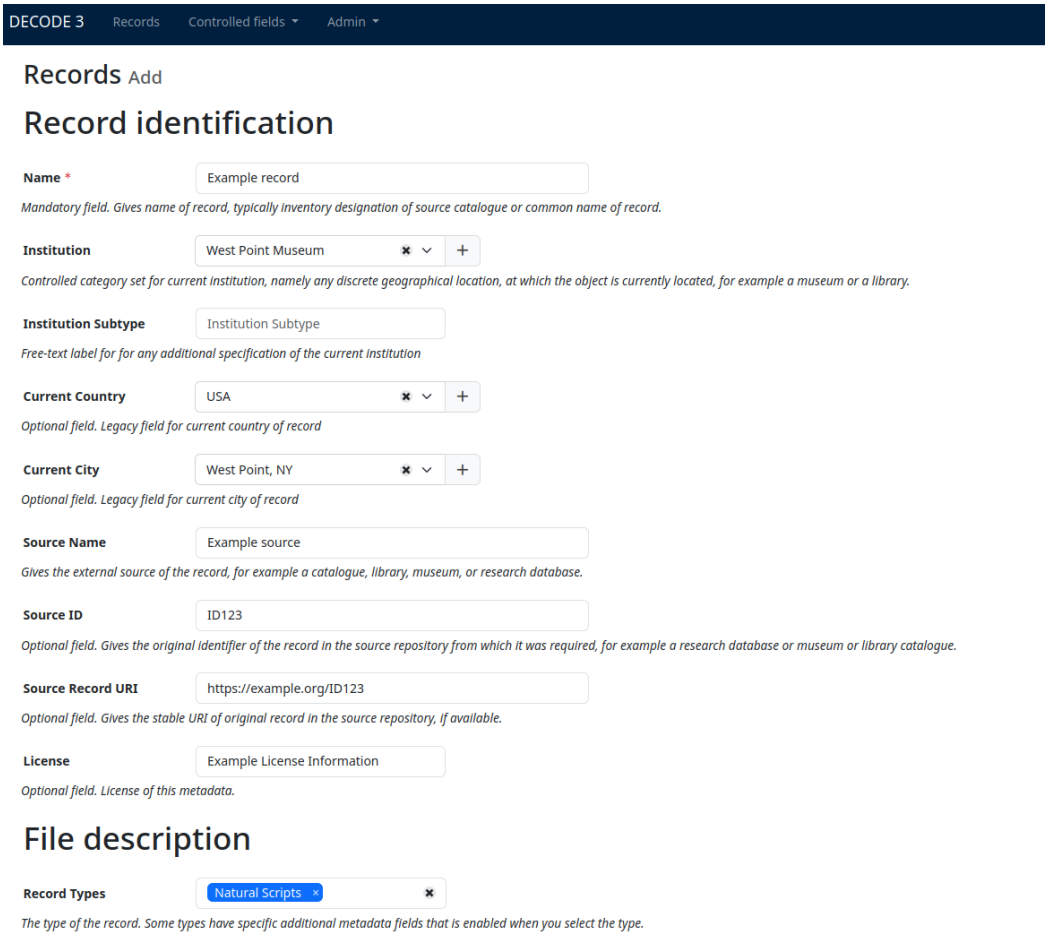

Figure 3: A screenshot of the DECODE database.

DESCRYPT is also intended to grow as a collaborative scholarly environment. The platform will support continued expert contribution through the addition of new records, revision of existing entries, and comparison of alternative readings and hypotheses. In parallel, the project aims to release stable data snapshots, benchmarks, and evaluation settings that can support transparent comparison of methods and encourage broader reuse within language technology and digital humanities.

In the longer term, DESCRYPT seeks to strengthen the visibility of rare writings within digital scholarship by making them more accessible as structured, analyzable, and reusable data.

7. Conclusion

This paper has introduced the design and implementation of a new corpus and database dedicated to rare and undeciphered scripts. The resource unifies heterogeneous materials—high-quality images, transcriptions, annotations, and metadata—within a standards-based infrastructure built for interoperability, reproducibility, and long-term preservation.

By combining established models from linguistic and cultural heritage data management with flexible solutions for low-resource and non-standard scripts, the database provides a foundation for systematic research across philology, linguistics, cryptology, and computer science. It supports both humanistic and computational workflows, enabling documentation, comparison, and analysis of writing systems that have long remained beyond the reach of large-scale digital methods.

Beyond corpus construction, the DESCRYPT project also aims to develop generic, data-lean pipelines for image processing, transcription, and symbol analysis, demonstrating how FAIR data principles and expert-in-the-loop methodologies can be applied to rare and complex scripts. Future work will expand coverage to underrepresented writing systems, enhance automated layout and symbol recognition, and deepen integration with infrastructures such as CLARIN and ELRA.

Ultimately, the initiative aims to transform rare scripts from isolated artifacts into analyzable linguistic data bridging cultural heritage preservation with computational approaches to the study of humanity's written past.

8. Ethical considerations

Building and sharing a corpus and database for rare and undeciphered scripts raises important ethical questions concerning cultural heritage, data stewardship, and responsible computational use. Many of the materials included in the resource originate from diverse cultural, religious, and political contexts, and some may carry particular significance for source communities or holding institutions. Digitization, description, and dissemination must therefore be undertaken with attention to provenance, attribution, rights, and, where relevant, appropriate restrictions on access.

A further ethical consideration concerns data ownership and reuse. Although the underlying sources are often historical, their digital reproductions, metadata, and annotations may involve new layers of intellectual contribution and institutional responsibility. DESCRYPT therefore seeks to balance openness and reuse with respect for licenses, archival agreements, and the interests of partner collections.

Responsible use of computational methods is equally essential. Automated analysis can support the study of rare writings, but it should not obscure uncertainty or replace expert judgement. For this reason, the project emphasizes transparency, traceability, and human oversight in transcription, annotation, and interpretation, so that computational outputs remain open to review, correction, and scholarly debate.

9. Limitations

The present resource remains a work in progress, and its coverage is necessarily uneven. Rare and undeciphered writings are often fragmentary, geographically dispersed, difficult to access, and documented according to highly variable scholarly conventions. As a result, some traditions are currently better represented than others, and the completeness and granularity of metadata, images, and transcriptions may vary across records.

A further limitation concerns standardization. Many of the scripts and symbol systems addressed in the project lack stable encoding conventions, established editorial practices, or agreed sign inventories. In such cases, provisional representations and working classifications are unavoidable, which may complicate comparison across sources and require later revision as scholarship advances.

The applicability of computational methods is also constrained by the nature of the material. Limited data availability, damaged or degraded surfaces, uncertain symbol boundaries, and unresolved linguistic affiliation all reduce the reliability of automatic approaches and make expert intervention indispensable. DESCRYPT therefore adopts an *AI-in-the-loop* approach, in which AI functions as an assistive tool within an expert-driven workflow: the scholar remains the principal agent of interpretation, while computational methods support analysis and hypothesis generation rather than driving the process and being corrected post hoc by the expert.

Finally, long-term sustainability depends on continued maintenance, technical development, and institutional support. Preserving the data, updating the platform, and adapting tools and workflows to changing research and infrastructure needs will require ongoing resources and active engagement from the wider scholarly community.

10. Acknowledgements

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: *Echoes of History: Analysis and Decipherment of Historical Writings* (DESCRYPT).

11. Bibliographical References

2024. Fabricius. Retrieved 2025-10-06.

2024. Transcribus. Retrieved 2025-10-07.

2025. eScriptorium. Retrieved 2025-10-06.

Nada Aldarrab. 2017. Decipherment of historical manuscripts. Master's thesis, University of Southern California.

Eugen Antal and Pavol Zajac. 2020. Hcportal overview. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, pages 18–20. Linköping Electronic Press.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

Ali Assarpour and Kent D. Boklan. 2010. How we broke the union code (148 years too late). *Cryptologia*, 34(3):200–210.

Federico Aurora. 2015. Damos (database of mycenaean at oslo). annotating a fragmentarily attested language. In Pedro A. Fuertes-Olivera et al., editor, *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, volume 198 of *Procedia - Social and Behavioral Sciences*, pages 21–31.

Alessandra Avanzini. 2009. Origin and classification of the ancient south arabian languages. *Journal of Semitic Studies*, 54(1):205–220.

Alison Babeu. 2011. *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*. Council of Library Information Resources.

Arnau Baró, Jialuo Chen, Alicia Fornés, and Beáta Megyesi. 2019. Towards a generic unsupervised method for transcription of encoded manuscripts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019)*, Brussels, Belgium.

Tim Berners-Lee. 2006. Linked data. `https://www.w3.org/DesignIssues/LinkedData.html`. W3C Design Issues.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data: The story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5:1–22.

Svenja Bonmann, Jakob Halfmann, Natalie Korobzow, and Bobomullo Bobomulloev. 2023. A partial decipherment of the unknown kushan script. *Transactions of the Philological Society*, 121:293–329.

John Chadwick. 1958. *The Decipherment of Linear B*. Cambridge University Press.

Jean-Francois Champollion. 1822. *Lettre à M. Dacier relative à l'alphabet des hiéroglyphes phonétiques*.

Jialuo Chen, Mohamed Ali Souibgui, Alicia Fornés, and Beáta Megyesi. 2020. A web-based interactive transcription tool for encrypted manuscripts. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, pages 52–59. Linköping Electronic Press.

Michael D. Coe. 2011. *Breaking the Maya Code*, 3rd edition. Thames & Hudson Ltd.

The TEI Consortium. 2023. *Guidelines for Electronic Text Encoding and Interchange P5 Version 4.7.0*. Last updated on 16th November 2023.

Shou de Lin and Kevin Knight. 2006. Discovering the linear writing order of a two-dimensional ancient hieroglyphic script. *Artificial Intelligence*, 170(4-5).

Francois Desset, Kambiz Tabibzadeh, Matthieu Kervran, Gian Pietro Basello, and Gianni Marchesi. 2022. The decipherment of Linear Elamite writing. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 112(1):11–60.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Empirical Methods in Natural Language Processing*.

Roger L. Easton and Keith Knox. 2016. Spectral imaging of manuscripts: Recovery of the past and preservation for the future. *Archiving Conference*, 2016(1):1–1.

Tom Elliotte, Gabriel Bodard, and Hugh Cayless. 2006–2022. Epidoc: Epigraphic documents in tei xml. Online material.

Bernhard Esslinger. 2024. *Learning and Experiencing Cryptography with CrypTool and SageMath*. Artech House, Norwood. Https://us.artechhouse.com/Learning-and-Experiencing-Cryptography-with-CrypTool-and-SageMath-P2378.aspx.

Silvia Ferrara and Fabio Tamburini. 2022. Advanced techniques for the decipherment of ancient scripts. *Lingue e linguaggio*, XXI(2):239–259.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Alica Fornés, Beáta Megyesi, and Joan Mas. 2017. Transcription of encoded manuscripts with image processing techniques. In *Proceedings of Digital Humanities*, Montreal, Canada.

Panorea Gaitanou, Ioanna Andreou, Miguel-Ángel Sicilia, and Emmanouel Garoufallou. 2022. Linked data for libraries: Creating a global knowledge space, a systematic literature review. *Journal of Information Science*, 50:204 – 244.

Geonames. GeoNames. Accessed: 2025-10-10.

Andrew George and Manfred Krebernik. 2022. Two remarkable vocabularies: Amorite-Akkadian bilinguals! *Revue d'assyriologie et d'archéologie orientale*, 116:113–166.

Getty Research Institute. 2025. Art & architecture thesaurus (AAT). Accessed: 2025-10-07.

Bradley Hauer and Grzegorz Kondrak. 2016. Decoding anagrammed texts written in an unknown language and script. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 75–86. MIT Press.

Mihály Héder and Beáta Megyesi. 2022. The decode database of historical ciphers and keys: Version 2. In *Proceedings of the 5th International Conference on Historical Cryptology, HistoCrypt 2022*.

Judith Hochberg, Patrick Kelly, Timothy Thomas, and Lila Kerns. 1997. Automatic script identification from document images using cluster-based templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):176–181.

Susan Hockey. 2008. The History of Humanities Computing. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, pages 3–19. Blackwell, Oxford & Malden, MA.

Uppsala universitet Institutionen för nordiska språk. 2020. Samnordisk runtextdatabas. http://www.nordiska.uu.se/forskn/samnord.htm. Accessed at 31 March 2026.

David Kahn. 1967. *The Codebreakers*, 2nd edition. New York.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2012. The copiale cipher. In *ACL Workshop on Building and Using Comparable Corpora (BUCC)*.

Nils Kopal. 2019. Cryptanalysis of homophonic substitution ciphers using simulated annealing with fixed temperature. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt 2019*, Mons, Belgium. NEALT Proceedings Series 37, Linköping Electronic Press.

Nils Kopal and Michelle Waldispühl. 2022. Deciphering three diplomatic letters sent by maximilian ii in 1575. *Cryptologia*, 46(2):103–127.

Benedek Láng and Beáta Megyesi. 2024. An STS analysis of a digital humanities collaboration: Trading zones, boundary objects, and interactional expertise in the DECRYPT project. *Humanities and Social Sciences Communications*, 11:618.

George Lasry. 2018. A methodology for the cryptanalysis of classical ciphers with search metaheuristics.

George Lasry, Norbert Biermann, and Satoshi Tomokiyo. 2023. Deciphering Mary Stuart's lost letters from 1578–1584. *Cryptologia*.

George Lasry, Beáta Megyesi, and Nils Kopal. 2020. Deciphering papal ciphers from the 16th to the 18th century. *Cryptologia*, pages 479–540.

Ernst Leierzopf, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021a. A massive machine-learning approach for classical cipher type detection using feature engineering.

Ernst Leierzopf, Vasily Mikhalev, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021b. Detection of classical cipher types with feature-learning approaches.

Peter Liddel. 2017. Greek inscriptions: insights and resources in the classroom and beyond. *Journal of Classics Teaching*.

Francisco Beltrán Lloris. 2015. The epigraphic habit in the roman world. In Ch. Bruun and J. Edmonson, editors, *The Oxford Handbook of Roman Epigraphy*, pages 131–148. Oxford University Press, Oxford—New York.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. volume 9, pages 69–81.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The DECODE database: Collection of historical ciphers and keys.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, Georg Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the decrypt project. *Cryptologia*.

Beáta Megyesi, Alicia Fornés, Nils Kopal, Benedek Láng, Michelle Waldispühl, Vasily Mikhalev, and Bernhard Esslinger. 2024a. Historical cryptology.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl de Leeuw, and Michelle Waldispühl. 2024b. Keys with nomenclatures in the early modern europe. *Cryptologia*, 48(2):97–139.

Cosimo Palma and Beáta Megyesi. 2025. DECODE2LOD: Connecting the DECODE Database with the Linked Open Data Cloud. In *HistoCrypt 2025 - The International Conference on Historical Cryptology*, Poznań.

Katerina Papavassileiou, Dimitrios Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *Journal on Computing and Cultural Heritage*, 16.

Eva Pettersson and Beáta Megyesi. 2018. The histcorp collection of historical corpora and resources. In *Proceedings of the Third Conference on Digital Humanities in the Nordic Countries*.

Pleiades. Pleiades: A Gazetteer of Past Places. Accessed: 2025-10-10.

Rune Rattenborg, Gustav Ryberg Smidt, Carolin Johansson, Nils Melin-Kronsell, and Seraina Nett. 2023. The archaeological distribution of the cuneiform corpus. *Altorientalische Forschungen*, 50(2):178–205.

Sujith Ravi and Kevin Knight. 2011a. Bayesian inference for Zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies (ACL-HLT 2011)*, pages 352–360, Portland, Oregon, USA. ACL.

Sujith Ravi and Kevin Knight. 2011b. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of ACL: Human Language Technologies (ACL-HLT 2011)*, pages 12–21, Portland, Oregon, USA. ACL.

READ-COOP SCE. 2024. Transkribus. Accessed: 2025-10-06.

Sravana Reddy and Kevin Knight. 2011. What we know about the voynich manuscript. In *ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Sravana Reddy and Kevin Knight. 2012. Decoding running key ciphers. *ACL*.

Annamaria De Santis and Irene Rossi. 2019. Crossing experiences in digital epigraphy: From practice to discipline. *De Gruyter*.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. *ACL*.

Isabel Velázquez Soriano and David Espinosa Espinosa. 2021. *Epigraphy in the Digital Age. Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions*. Archaeopress Archaeology, Oxford.

Mohamed Ali Souibgui, Asma Bensalah, Jialuo Chen, Alicia Fornés, and Michelle Waldispühl. 2022. A user perspective on HTR methods for the automatic transcription of rare scripts. the case of Codex Runicus. *ACM Journal on Computing and Cultural Heritage*.

Michael P. Streck. 2010. Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus. *Mitteilungen der Deutschen Orient-Gesellschaft*, 142:35–58.

Ferenc Szigeti and Mihály Héder. 2022. The transcript tool for historical ciphers by the decrypt project. In *Proceedings of the 5th International Conference on Historical Cryptology*, pages 208–211.

Melissa Terras. 2012. Image processing and digital humanities. In N. M. Terras, J. Nyhan, and C. Warwich, editors, *Digital Humanities in Practice*. Facet Publishing.

Melissa Terras. 2016. *Crowdsourcing in the Digital Humanities*. Wiley-Blackwell.

The Unicode Consortium. 2019. *The Unicode® Standard, Version 12.0 – Core Specification*. The Unicode Consortium, Mountain View, CA.

Thesaurus Linguae Aegyptiae. 2025. Thesaurus Linguae Aegyptiae. Accessed: 2025-10-08.

Trismegistos. 2025. Trismegistos. https://www-trismegistos-org. Accessed: 2025-10-08.

Wikimedia Foundation. 2025. Wikidata: A free and open knowledge base. Accessed: 2025-10-07.

Henrik Williams, Marco Bianchi, and Christiane Zimmermann. 2022. Corpus editions of runic inscriptions in supranational databases. *Futhark: International Journal of Runic Studies*, 12:117–135.

Xusen Yin, Nada Aldarrab, Beáta Megyesi, and Kevin Knight. 2019. Decipherment of historical manuscript images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 78–85, Sydney, NSW, Australia.

12. Language Resource References

Federico Aurora. 2015. DAMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. In Pedro A. Fuertes-Olivera et al., editor, *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, volume 198 of *Procedia - Social and Behavioral Sciences*, pages 21–31.

Uppsala universitet Institutionen för nordiska språk. 2020. Samnordisk runtextdatabas. `http://www.nordiska.uu.se/forskn/samnord.htm`. Accessed at 31 March 2026.

Beáta Megyesi and Mihály Héder and Benedek Láng. 2024. *The DECODE database: A corpus of historical ciphertexts and cipher keys*. Department of Linguistics, Stockholm University. PID de-crypt.org. Retrieved 2025-10-06.

Eva Pettersson and Beáta Megyesi. 2024. *HistCorp: Historical Corpora*. Department of Linguistics and Philology. PID https://www2.lingfil.uu.se/person/pettersson/histcorp/. Retrieved 2025-10-06.