# The DECODE Database of Historical Ciphers and Keys: Version 2

**Mihály Héder**
Dept. for Philosophy and History of Science
Budapest University of Technology and Economics;
ELKH SZTAKI
Hungary

**Beáta Megyesi**
Dept. of Linguistics and Philology
Uppsala University
Sweden

## Abstract

We report recent developments of the DECODE database aimed for the systematic collection and annotation of encrypted sources: ciphertexts, keys and related documents. We released a new, more functional graphical user interface, revised some metadata features and enlarged the collection and tripled its size.

## 1 Introduction

In recent years we have seen an increased interest in developing infrastructural resources for historical cryptology, like various types of cipher collections and tools to break them. Among the resources, we can mention websites about single ciphers such as the excellent portal about the Voynich manuscript by René Zandbergen[1] for the Voynich lovers, or Klaus Schmech's popular blog pages[2] about many different encrypted sources. There are also online resources aimed at a more or less systematic collection of historical encrypted manuscripts. One is Satoshi Tomokiyo's private website `Cryptiana`[3] which contains an ambitious collection of historical ciphertexts and keys from various libraries and archives. Another resource currently under development is the `Portal of Historical Ciphers`[4] by Eugen Antal and Pavol Zajac (Antal and Zajac, 2020) which is a database for historical encrypted manuscripts released with a user-friendly graphical interface (GUI).

The DECODE database (Megyesi et al., 2019) which we will focus on in this paper was developed with the aim to serve as a long-term storage for the collection and description of historical encrypted sources. The very first release of a development version took place in 2016 with a few hundred manuscripts made available to the public, along with a description of the whereabouts of the sources. The first fully developed version of the database was released and published in 2019 (Megyesi et al., 2019) with nearly thousand records.

In this paper, we describe the new version of the database[5], present its new design and search interface, the new metadata features, as well as the new, enlarged collection which tripled in size during the past two years.

## 2 DECODE: Version 2

The initial version of DECODE facilitated the collection of material very well. However, over the two years enough user feedback was collected for a redesign, both for user-facing features and the underlying software. Finally, the creation of a new version was also necessary to facilitate the needs of the *DECRYPT* project (Megyesi et al., 2020), that aims to develop an integrated, web-based pipeline for for the automatic transcription and decryption of ciphertexts to historians and others with interest in historical cryptology. While the changes in the software architecture are out of scope of this paper, we present the most relevant changes in the user interface.

### 2.1 Interface

One of the most popular features is the search function. Therefore, we paid special attention to polish this functionality. In DECODE v2, the user is presented with simple search right on top of the record list that allows for keyword-based look-ups, as well as an advanced search, that used to rely on logical operators. Instead, a GUI with filtering

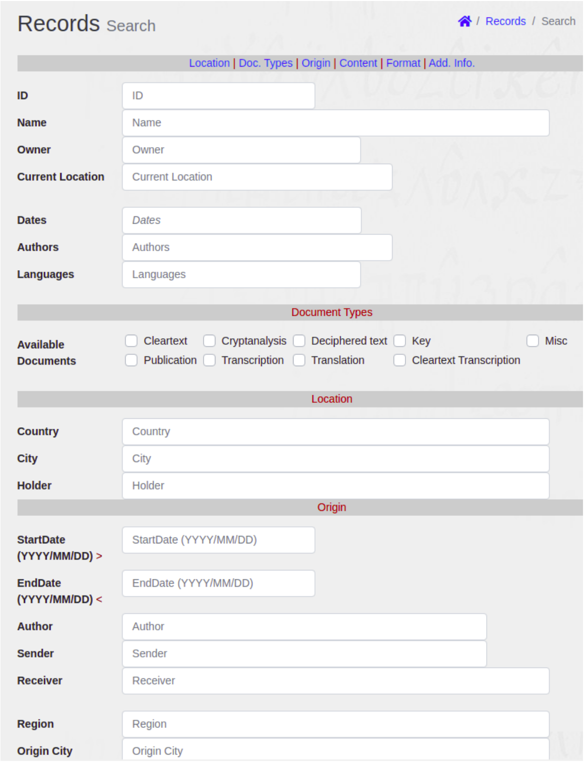
[1] http://www.voynich.nu/

[2] https://scienceblogs.de/klausis-krypto-kolumne/

[3] http://cryptiana.web.fc2.com/code/crypto.htm

[4] https://hcportal.eu/

[5] The service is available at `https://de-crypt.org/decode`.



Records Search

Location | Doc. Types | Origin | Content | Format | Add. Info.

| | |
|---|---|
| ID | ID |
| Name | Name |
| Owner | Owner |
| Current Location | Current Location |
| Dates | Dates |
| Authors | Authors |
| Languages | Languages |

Document Types

Available Documents
□ Cleartext □ Cryptanalysis □ Deciphered text □ Key □ Misc
□ Publication □ Transcription □ Translation □ Cleartext Transcription

Location

| | |
|---|---|
| Country | Country |
| City | City |
| Holder | Holder |

Origin

| | |
|---|---|
| StartDate (YYYY/MM/DD) > | StartDate (YYYY/MM/DD) |
| EndDate (YYYY/MM/DD) < | EndDate (YYYY/MM/DD) |
| Author | Author |
| Sender | Sender |
| Receiver | Receiver |
| Region | Region |
| Origin City | Origin City |

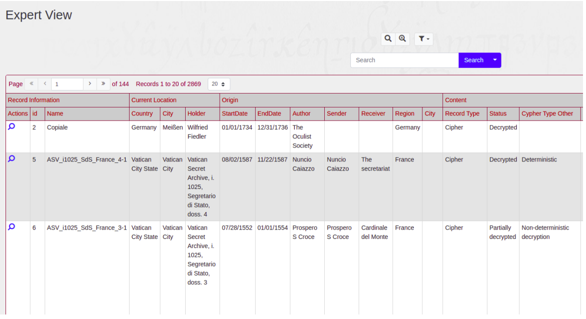
Figure 1: Some filtering options available in advanced search.

capabilities, that is more familiar to most users is implemented, and illustrated in Figure 1.

Another commonly used feature is for advanced users. In a screen we now call ”expert view” the users can browse the data with all details of the metadata presented at once, as shown in Figure 2. In addition, they have the capability of exporting their search results in various commonly used formats such as Excel, HTML, and CSV allowing comma-separated values.

The envisaged integration with other tools in the tool-chain aimed for the analysis of encrypted sources is best served by an API[6] that provides all the features that are also available on the web. By using this REST API not only our tools may be integrated but our users may come up with their own workflows.

A highly practical enhancement for the users who are collecting and uploading records is the possibility of mass-upload of scanned documents. This feature allows the transfer and processing of 100s of documents at once in a partially parallellized manner, leveraging modern web technologies.



Expert View

Page « ‹ 1 › » of 144 Records 1 to 20 of 2869 20

| Record Information | | | Current Location | | | Origin | | | | | | | Content | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actions | id | Name | Country | City | Holder | StartDate | EndDate | Author | Sender | Receiver | Region | City | Record Type | Status | Cypher Type Other |
| 🔍 | 2 | Copiale | Germany | Meißen | Wilfried Fiedler | 01/01/1734 | 12/31/1736 | The Oculist Society | | | Germany | | Cipher | Decrypted | |
| 🔍 | 5 | ASV_i1025_SdS_France_4-1 | Vatican City State | Vatican City | Vatican Secret Archive, i. 1025, Segretario di Stato, doss. 4 | 08/02/1587 | 11/22/1587 | Nuncio Caiazzo | Nuncio Caiazzo | The secretariat | France | | Cipher | Decrypted | Deterministic |
| 🔍 | 6 | ASV_i1025_SdS_France_3-1 | Vatican City State | Vatican City | Vatican Secret Archive, i. 1025, Segretario di Stato, doss. 3 | 07/28/1552 | 01/01/1554 | Prospero S Croce | Prospero S Croce | Cardinale del Monte | France | | Cipher | Partially decrypted | Non-deterministic decryption |

Figure 2: A portion of the expert view.

Finally, an important non-functional property of the system is its improved speed. While we did not make measurements about the old system — nor is the underlying hardware completely similar — but based on the user feedback, haste of the execution is one of the most popular improvements of DECODE v2.

## 2.2 Metadata

Apart from interface improvements, we also made some additions to the structure of metadata. These changes are gradual and small, reflecting our cautious approach to modifying the data schema.

One change enacted is a finer distinction in the system between the creator of a record and the owner of a record; this is in anticipation of a division of labour between uploaders and curators.

A better handling of the dates is enabled by the separation of *year*, *month* and *day of month*, which makes it possible to mark any subset of the three as unknown, instead of forcing the users to set a given complete date or leave the date out.

Also, DECODE v2 facilitates the academic collaboration by enabling the users to add publications to any given records — that is, academic publications that discuss the historic record in question.

Finally, an improved handling of cross-record relationships is also enabled, such as allowing the connection and linking between related key(s) and ciphertext(s). Currently, this relationship is for key–cipher pairs, but in the future we envision other kinds of relationships.

# 3 Collection

The database contains a collection with the largest number of historical encrypted sources of today. As of April 4, 2022, 2 939 records have been uploaded out of which 1 272 (44%) are ciphertexts and 1 667 (56%) cipher keys. They come from

[6] https://de-crypt.org/decrypt-web/swagger/

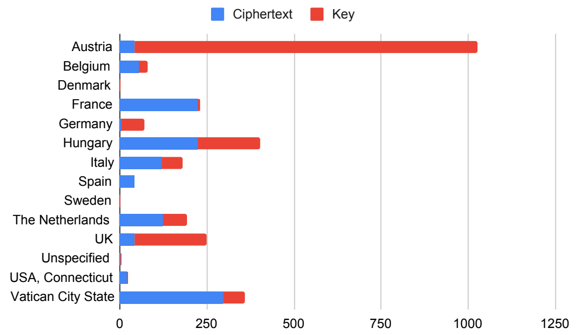


Figure 3: The distribution of ciphertexts and keys across holder countries in the DECODE database.

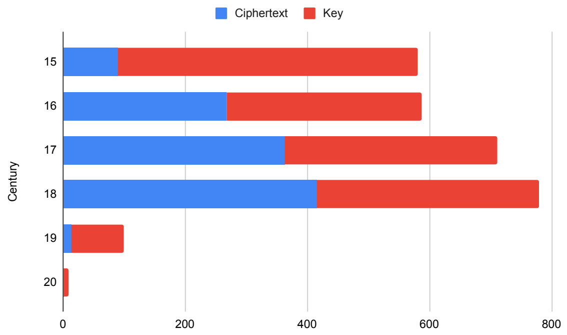


Figure 4: The distribution of ciphertexts and keys across centuries in the DECODE database.

archives and libraries from 13 countries: Austria, Belgium, Denmark, France, Germany, Hungary, Italy, Spain, Sweden, the Netherlands, UK, USA, and the Vatican City State, as shown in Figure 3. The number and type of records differ greatly across the holder countries. Most records come from Austria, followed by Hungary, the Vatican, the UK, Italy, and the Netherlands, and the fewest with only a handful records come from the Nordic countries. The reason to the uneven distribution is because of our collection strategies. We search for and vacuum-clean encrypted sources archive by archive and try to do it systematically. The collection of sources is usually carried out by a historian in collaboration with a librarian from a particular archive, and the upload is done either by the historian or by a trained research assistant to enter the correct metadata fields. We get also in touch with researcher and other interested about single ciphers. If the source is proved to be original with some kind of documentation and information about its whereabouts, we upload the source as well. For example, we have a few ciphers and keys from the Nordic countries (Denmark and Sweden) although we have not searched for encrypted sources in those countries as yet.

The collection contains sources originating mainly from Early Modern times (15th-18th centuries) but there are also ciphers and keys from later periods. The distribution of encrypted sources between the 15ht and 18th centuries are quite even, but the type of sources are less balanced as we have access to more keys from the 15th and 16th centuries than ciphertexts, as also visualized in Figure 4.

The sources consist of in total 18 788 images of 128 GB in total, which gives 6.8 MB/image on average. Most of the sources are one-page long but there are usually additional pages to each record containing related pages with signatures, dating and any other information about the sender/receiver of the message.

Out of all records, 1 111 have been manually transcribed, following the conventions and guidelines developed for a consistent transcription across ciphers and keys (Megyesi, 2020; Megyesi and Tudor, 2021).

## 4 Future Plans

Our ambition and long-term goal is that the database would be fully functional and integrated with transcription and decryption tools and would serve as the first, natural place to look for encrypted sources with adequate and consistent metadata and cryptanalysis. To achieve this goal, we need to work on the implementation of integration of tools for transcription and decryption and carry out user studies to improve the search interface.

A main theme of the immediate future is to integrate various tools of analysis in our system. This includes a transcription tool, various cryptanalysis tools and tools to analyze the linguistic structure and the nomenclatures of the keys, along with the inclusion of historical language models in transcription and decryption.

A major part of this work is to enable our users to work with a document separately from the official record; that is to enable them to conduct trials and experiments on the records and propose new results to the curators of the given record. This way several plaintext suggestions may be proposed by different groups or individuals for any given ciphertext record. These interactions come with substantial technological challenges. The machine learning facilitated image processing algorithms typically require the usage of GPUs that in turn require the implementation of careful resource management, as the common graphical processing units cannot be time-shared as of today. Another key area of work is the data model required for the collaboration of tools created by various parties — another task for which the DECRYPT project is well placed.

In the database, we will increase the set of metadata for describing cipher keys to analyze the plaintext, the ciphertext and the encoding structure in keys with and without nomenclatures. Furthermore, we would like to extend the information about physical person and geographic areas in the metadata to be able to use Linked Open Data to connect historical persons and areas. Also, we will carry out more consistency checks across features, such as the description of dating and place of use where such information is missing.

We will also continue increasing the collection of historical ciphers and keys from various archives and libraries, mainly from Europe. We would like to encourage all our current and future users to not hesitate to reach out to us to upload new ciphers or edit existing ones thereby contributing to our field. We can provide help with the time-consuming uploads to fill in metadata and connect related documents. We do hope that the resource will further strengthen the research in historical cryptology which also allows for some fun for code-breakers.

## 5 Conclusion

In this paper, we describe the recent developments of the DECODE database allowing storage, search and analysis of encrypted sources. We presented a new graphical user interface with improved, more intuitive and fast advanced search functions, export of data in excel, xml or CSV formats, improved metadata to search for dates and information about the owner and contributor to the record. We increased the collection and tripled its size in 3 years to make available almost 3 000 ciphertexts and keys to the users. The new version of the DECODE database with open API, fast search and more intuitive user interface provides the basis for the integration to a pipeline for (semi-)automatic decryption of ciphers.

## Acknowledgments

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts. We would like to thank our colleagues in the DECRYPT team, in particular Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, George Lasry, Karl de Leeuw, Anna Lehofer, Eva Pettersson and Crina Tudor for their input, suggestions on the design, and contributions. We would like to send our appreciation to Anne-Simone Rous and Satoshi Tomokiyo for generously sharing their resources with all of us.

## References

Eugen Antal and Pavol Zajac. 2020. HCPortal overview. In *In Proceedings of the 3rd International Conference on Historical Cryptology (HistoCrypt 2020)*, pages 18–20.

Beáta Megyesi and Crina Tudor. 2021. Transcription of Historical Ciphers and Keys. Guidelines, version 2.0. `https://cl.lingfil.uu.se/~bea/publ/transcription-guidelines-v2.pdf`. Version: March 30, 2021.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The DECODE Database: Collection of Ciphers and Keys. In *Proceedings of the 2nd International Conference on Historical Cryptology (HistoCrypt 2019)*, Mons, Belgium, June.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, George Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the DECRYPT project. *Cryptologia*, 0(0):1–15.

Beáta Megyesi. 2020. Transcription of Historical Ciphers and Keys. In *Proceedings of the 3rd International Conference on Historical Cryptology (HistoCrypt 2020)*, Budapest, Hungary, June.