The DECODE Database
Collection of Historical Ciphers and Keys

Beáta Megyesi, Nils Blomqvist and Eva Pettersson
Department of Linguistics and Philology
Uppsala University, Sweden
firstname.lastname@lingfil.uu.se

Abstract

We present an on-line database DECODE consisting of encrypted historical manuscripts, aiming at the systematic collection of ciphers and keys to create infrastructural support for historical research in general, and historical cryptology in particular. The collected material is annotated with a metadata scheme developed specifically for historical ciphers. Information includes provenance and location of the manuscript, computer-readable transcription, possible decryption(s) of the ciphertext and translation(s) of the plaintext, images, and any additional materials of relevance to the particular manuscript. The database allows search in the existing collection and upload of new encrypted sources by users.

1 Introduction

According to some historians, 1% of the national archives and libraries in Europe contain secret messages, encrypted hand-written manuscripts, intended to hide the content of the message with the exception of the intended receiver(s). Keys used for encryption might also be found, often without being stored together with the encrypted or decrypted message. The manuscripts in the libraries and archives are seldom indexed as ciphers, which makes it difficult to find them unless you know the librarian with extensive knowledge about the library's selection. Historians and other scholars interested in our history stumble on these manuscripts when searching for sources from a particular period of their interest. They try to decrypt the hidden source to shed some new light on our history — to find new interpretations and explanations.

However, it is far from trivial how to crack a secret message and there is a clear lack of infrastructural support in terms of data resources and automatic tools that historians and others without any knowledge in cryptology can use to reveal the content of the hidden message. In order to develop tools for automatic decryption, we need large(r) collections of ciphertexts and keys to develop better algorithms and systematically evaluate them on various cipher types.

In this paper, we describe an on-line database, DECODE[1] aiming at the systematic collection and description of ciphers, keys and related documents. The database comes with a graphical user interface that allows simple and advanced search in the existing collection for all users, and upload of new ciphertexts and keys by users with an account.

The DECODE database is one of the first steps towards an infrastructure for historical cryptology. Our goal is to collect ciphertexts, codes, keys, and codebooks from various archives and libraries as well as from the public, and to develop tools to support the transliteration of images into computer-readable format, cryptanalysis, and to make the resources and tools available to people interested in historical cryptology. Our hope is that users will contribute to enlarge the database by uploading new material for a growing collection, a monitor corpus of historical ciphers and keys.

The paper is structured as follows. In Section 2, we describe the general architecture of the database followed by a detailed description of the metadata, a set of features with their possible values for the description of the encrypted manuscripts in Section 3. Information

[1] https://cl.lingfil.uu.se/decode

Proceedings of the 2nd International Conference on Historical Cryptology, pages 69-78
Mons, Belgium, 23-26 June, 2019

about cryptanalysis and decryption, including standards for transcription/transliteration of the images is described in Section 4. Then, we present the search design in Section 5, and the upload and editing functions in Sections 6 and 7. User access roles are described in Section 8 followed by a brief technical description of the database in Section 9 and some tools for the automatic processing of ciphers in Section 10. Lastly, in Section 11, we conclude the paper and give directions for future research.

2 The DECODE Database

The database was developed between 2015-2018, to cover a large range of different cipher types and keys for various plaintext languages from early modern time. We expected that the Vatican archives would be the right place for our endeavor, given papal correspondence throughout the centuries with many countries, with many (European) languages. Luckily, the Secret archives of the Vatican and the main library of the Vatican are well-organized archives with indexes over the encrypted manuscripts' whereabouts and description of their provenance. Within a few weeks, we could collect over 300 ciphers and some keys, and order images from the archive. This dataset became the starting point i) for a systematic description of ciphers and keys, ii) to develop the database with a search function, iii) to draw up guidelines for transliteration of ciphers, iv) to develop tools for semi-automatic transcription using hand-written text recognition, v) to implement tools for cryptanalysis, and vi) to map available ciphers with their corresponding key in the database.

During the past year, the database has been publicly released and three historians were asked to upload their cipher and key collection to test the functionality of the database. At the time of writing, the collection contains nearly 1000 records, ciphers and keys with images, and description of their current location, provenance, content, along with related documents including transcriptions/transliterations, cryptanalysis, related generated key, deciphered plaintext, possible translation(s), references to publications and other information of interest. Each record is marked with the name of the user who uploaded the record and the date for entering the record into the database.

The ciphers and keys are collected at various archives and libraries in Austria, Belgium, Germany, Hungary, Italy, the Netherlands, UK, and the Vatican City. Among the dated records, the earliest ones originate from the 15th century, and the latest from 1793. About 33% of the material consists of original keys. Out of 634 ciphers, 205 are decrypted, and 232 are transcribed as running text allowing further processing for cryptanalysis. Among the records, we find plaintext languages in Dutch, English, French, German, Hungarian, Italian, Latin, or a combination of these (e.g. English-Latin, Hungarian-Latin, Italian-Spanish).

The majority of the records are short, one-page images, but we also find longer ciphers, the longest 410 pages. The great majority of the ciphers are encrypted with numbers, but ciphers with alphabetic characters and esoteric symbols such as zodiac and alchemical signs are also present. The known cipher types in the database are mostly based on simple substitution or homophonic substitution, with or without nomenclatures, but polyphonic substitutions also appear.

To browse the database, we developed simple and advanced search functions that are available to the public. The graphical interface is accessed using a web browser, making it usable on various operating systems and devices (smartphones, tablets, etc.) Experts in the field of historical cryptology may also apply for a database account to be allowed to add new records, and edit existing ones. In order to be able to enter and store a record in the database, image(s) showing the original cipher or key are required to ensure that the record actually exists. Images and other related documents that cannot be distributed for copyright reasons can be marked as private by the user. In such cases, the private documents/images are accessible only to the owner of the record, i.e. the person who uploaded the original document, and the database system owner(s). The database is publicly available for search but private images or documents related to the records are neither visible nor accessible to anyone, except for the owner. For

users with login, private images/documents are shown as miniature pictures but the documents are not downloadable.

On the basis of the initially collected records present in the database, we developed a metadata scheme for the description of ciphers and keys, which will be described next. Our hope is that the metadata can serve as basis for a standardized description of ciphers and keys.

3 Describing Ciphers: Metadata

Each manuscript is described according to a subset of metadata structured as attribute-value pairs. The structured information is divided into three fields: the current location of the manuscript, information about its content, and the form of the manuscript.

3.1 Current Location

The current location of the manuscript, being it a ciphertext or a key, is mandatory information divided into three attributes: *Country*, *City*, and *Holder*. Country and City relate to the current location of the document, while Holder refers to an institution or a person who owns or keeps the document today.

3.2 Origin

The field *Origin* gives information about the origin of the manuscript: the *Dating* or time period when the document was created, the name of the *Author* of the manuscript, the *Sender*(s) and the *Receiver*(s), which can be an institution or a person, as well as the place, the *Region* and the *City* of origin. Given that we often do not know much about the provenance of the manuscript, information about the origin of the document is supplemental. However, knowing something about the provenance of the manuscript might be highly valuable during the puzzle of the decryption process, for example to make educated guesses about the possible plaintext language(s) behind the encrypted document.

Dating can be given as a specific date or a time period during which the document is assumed to have been created. Dates are represented in the proleptic Gregorian calendar, and follow the convention of ISO 8601 where 1 BCE is represented as year 0, 2 BCE is represented as -1, and so on. The year, month and day (in that order) must be separated by a dash (-). The user can enter the year only, and it is interpreted as an interval spanning over the entire year. For example, 1542 is interpreted as an interval between 1542-01-01 to 1542-12-31 (i.e. from January 1, 1542 to December 31, 1542). If the exact dating of the document is not known, an interval can be specified with a starting and ending date, separated by a colon (:). Thus, 1542:1570 is interpreted as the interval between 1542-01-01 to 1570-12-31. If the month or day is not specified in the ending date, they will default to the last day of the last month. For example, 1542-05:1570-03 is interpreted as the interval between 1542-05-01 to 1570-03-31.

Concerning naming the sender and receiver of the manuscript, and naming the city or region where the manuscript originates from, the user is allowed to fill in the full or partial name of the *Author*, the *Sender*(s) and the *Receiver*(s). The *Region* and *City* of origin might be given as original names, i.e. the historical name of the region or city used when the document was created. For example, if a manuscript originates from the current capital of Slovakia, Bratislava, it is up to the user to decide whether to name the city by its old German name *Pressburg* , use the old Hungarian name *Pozsony*, or give the current name *Bratislava*. We are aware of the fact that the lack of standard concerning the name of cities and regions throughout the history might confuse users, and raises higher demands on the user when searching in the database.

3.3 Content

Describing the content of the encrypted document includes ten different types of optional attributes. *Number of pages* gives, as its name indicates, the number of pages of the ciphertexts or key given the image, excluding entirely non-encrypted plaintext pages, such as title pages or envelope. This is intended for quick search, for example, for shorter or longer ciphertexts.

The manuscript can either be a cipher, or a key, which is defined in the *Cipher/Key* field as predefined values. Each cipher can also be marked on the basis of its *Status*, in terms of whether the cipher is decrypted, non-decrypted, or partly decrypted. For keys, the

value of *Status* is non-applicable (N/A). For each record, it is indicated if it is a publicly available record with downloadable images and documents, i.e. if the record is *Public cipher/key* or not.

Cipher type is predefined and can be added to keys, ciphers and codes. If the type is not known, the value UNKNOWN applies. In case the type is known, the following predefined values might apply: homophonic substitution, nomenclatures, polyalphabetic, simple substitution, transposition, or a combination of those. If the cipher is of another type than the predefined values, there is an option to fill in a user-defined one in the OTHER field.

Encrypted documents consist of many different *Symbol sets*, numbers, alphabetical characters from various alphabets (e.g. Latin or Greek letters), diacritics, esoteric symbols such as zodiac or alchemical signs, logograms, punctuation marks, and a combination of those. In the database, the user is allowed to give information about the symbol set used in a particular manuscript, predefined values for alphabet, esoteric, and/or numerical symbols. White space in the original ciphertext might be present intentionally to keep word boundaries during encryption, between code groups, or unintentionally when producing the encrypted message. The user can indicate whether the ciphertext includes white space or not. If the encrypted manuscript contains other symbols than the pre-defined list of symbols, the user can define his/her own.

Lastly, the encrypted manuscript might contain encrypted sequences, i.e. ciphertext only, but plaintext, non-encrypted sequences of words, sentences, paragraphs and even pages might occur embedded inline in the message. Figure 1 shows an encrypted message with the plaintext <comé la mi comanda> in the ciphertext (ASV, 2016a).

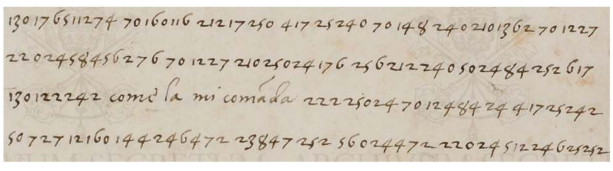

Figure 1: Extract from a cipher, with cleartext embedded in the ciphertext.

The cipher might contain not only embedded non-encrypted plaintext, but also decrypted plaintext, often written over the ciphertext sequences by the receiver, see Figure 2 (ASV, 2016b).

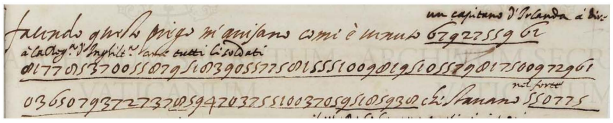

Figure 2: Extract from a cipher with cleartext embedded in the ciphertext, and decrypted plaintext.

Therefore, there is a need to differentiate between the plaintext, written in the original language and in which the ciphertext is embedded (or vice versa), and plaintext representing the decrypted ciphertext. We define plaintext as the decrypted ciphertext, and cleartext, as a non-encrypted text embedded in the message. According to the above, we indicate in the metadata description whether the text contains *Inline cleartext*, *Inline plaintext*, or both. In Figure 2, both *Inline cleartext* and *Inline plaintext* are added as values.

Similarly, the *Cleartext language(s)* (if any) and the original underlying language of the cipher, i.e. the *Plaintext language* can be defined by the user, as optional fields.

3.4 Format

Another optional field is the *Format* aiming at the description of the *paper* or *ink type*, for codicological studies to date a particular manuscript. The user can fill in these fields as free text. Typically, these fields are used for notes about the quality of the paper/parchment, whether it is damaged, or difficult to read or interpret due to bleed-through ink, or just a bad photocopy.

3.5 Other

There is also an option to provide links to available publications and other information about the manuscript in text format in the field *Additional information.*

3.6 Related Documents

In addition to the description of the ciphertext or key given as metadata fields, there is the possibility to add various types of documents describing the manuscript. These can

be transcriptions or transliterations of the manuscript (*Transcription*), the decrypted or decoded plaintext of the cipher (*Deciphered text*), generated key(s) (*Key*), statistics and cryptanalysis of the ciphertext (*Cryptanalysis*), or any other relevant document (*Miscellaneous*).

Many manuscripts containing ciphertexts or keys are buried among letter correspondences written in cleartext. These might be of relevance for the historical interpretation and contextualisation of the manuscript and could also be helpful for cryptanalysis to make educated guesses about the topic of the document, to crack encoded named entities such as place or personal names, and so on.

The plaintext might be translated and here, the user can upload *Translation*(s) of the plaintext to various languages.

If there is any published material about the cipher or key, the user can upload these (*Publications*), or add references as a single file.

The documents can be uploaded in various formats (txt, doc, docx, pdf) and most image types are allowed (e.g. png, tiff, jpeg). When uploading a document, the user is asked to name the document and categorize it by its type. The user can decide whether to create the documents with free access and downloadable to everyone, or to keep these private so other users won't be allowed to access those. The images can be uploaded as a single file, or as multiple files stored in the same folder by holding the Control key (Windows/Linux) or the Command key (Mac) while clicking on the desired files.

4 Analyzing Ciphers

In addition to the cipher collection and browsing function provided, we develop tools for the automatic transcription and decryption of ciphers. Given an image representing a cipher or key, the first step is to transcribe or transliterate the ciphertext into a computer-readable format. Then, the transcribed ciphertext can be statistically analyzed using various metrics (n-gram frequency, clusters, index of coincidence, entropy measures), and decrypted. Keys and ciphers can also be mapped automatically, calling for language models for various European languages. In the subsequent sections, we give an overview of the transcription guidelines, and tools for cryptanalysis and cipher-key mapping.

4.1 Transcription/Transliteration

Usually, the first step in attacking a cipher is the conversion of the image into a machine-readable format, represented as text. There are many different ways of transcribing or transliterating a manuscript. Therefore, we developed guidelines so that the transcriptions available in the database have a common format.

Each transcript file of a particular cipher (which may consist of multiple images) starts with comment lines with information about the file. Each comment line starts with "#" followed by a transcription attribute and its value, as illustrated below:

- #CATALOG NAME: your own index, i.e. file location: e.g. /Segr. Stato Francia 6/1
- #IMAGE NAME: the name of the image(s) representing the cipher: e.g. image interval 234r-237v.jpg
- #TRANSCRIBER NAME: full name or initials of the transcriber: e.g. BeMeg
- #DATE OF TRANSCRIPTION: the date the transcription was submitted
- #TRANSCRIPTION TIME: the time it took to transcribe all images of a cipher in hours and minutes without counting breaks and quality checks
- #COMMENTS: description of e.g. difficulties, problems

Next, the content of the image is transcribed. Each new image in a cipher starts with a new comment line with information about the name of the image followed by a possible comment line:

- #IMAGE NAME: the name of the image, e.g. 234v.jpg
- #COMMENTS: any comments, e.g. difficult to read line 3, bleed-through

The transcription is carried out symbol by symbol and row by row keeping line breaks,

spaces, punctuation marks, dots, underlined symbols, and cleartext words, phrases, sentences, paragraphs, as shown in the original image. Line breaks are kept so that when a new line starts, a new line is added in the transcription. Punctuation marks, such as periods, commas, and question marks are transcribed as such. Space is represented as space. If there is a larger width of a space in relation to other spaces in the ciphertext, two or more space characters can be entered in the transcription. The reason for allowing several space characters is that a larger space in the original might mark word boundaries which the encryptor unintentionally left there when encrypting the manuscript, which can be helpful in the decryption process as they might denote word boundaries.

Sometimes, punctuation marks (e.g. dots, commas, underscores) appear above or under specific symbols. It could be ink splash, but if they appear in a systematic way, they are transcribed as well. If the mark appears above the symbol, the sequence is transcribed as the symbol, followed by "∧" and the specific mark (e.g. dot or comma). If the mark appears under the symbol, it is marked by an"_" placed between the symbol and the mark "." (e.g. s_.). Similarly, underlined symbols are marked with "_" immediately following the symbol, except when the whole ciphertext is underlined.

Uncertain symbols are transcribed with added question mark "?" immediately following the uncertain symbol. Possible interpretations of a symbol can be transliterated by transcribing the options using the delimiter "/". For example, if it is not clear if a symbol represents a 0 or 6, it is transcribed as "0/6?".

The cipher sequences might be embedded in cleartext, or cleartext might be embedded in ciphertext, see Figure 3. We can also find cleartext in keys, often explanations about the key. To be able to distinguish between ciphertext and cleartext sequences, the latter is clearly marked in brackets as <CLEARTEXT LANG letter/word sequence>, where the tag <CLEARTEXT ... > denotes where the cleartext starts and ends. LANG represents the language the cleartext is written in, marked by ID as defined by ISO 639-1 two-letter codes for languages (e.g. IT for Italian), and UN for unidentified languages.

Figure 3: Transcription of the cipher image in Figure 1.

Sometimes we can find the decrypted plaintext written above the ciphertext. Similarly to cleartext, plaintext is transcribed as <PLAINTEXT LANG letter/word sequence> in a separate line. Transcription of the image containing cleartext and plaintext of the original image in Figure 2 is shown in Figure 4. Note that if the cipher is cracked, the entirely or partly deciphered parts, i.e. the plaintext, can be uploaded as a separate file (see Section 3.6).

Transcription reflects the intention of the encoder, i.e. the corrected segments are transcribed. For example, if numbers are crossed-off in the original, these are not transcribed. Similarly, insertions of corrections between symbols are transcribed, as they intended to appear in the original. Ciphertext/cleartext written in the margin is added into the specific space as indicated by the given mark/note in the original cipher.

Not seldom, historical manuscripts contain catchwords placed at the foot of the page to mark page order (instead of numbers). Catchwords are a sequence of symbols anticipated as the first symbols of the following page. In ciphers, catchwords might denote an actual word, unintentionally, and therefor transcribed as <CATCHWORD symbol sequence> (e.g. <CATCHWORD 1 1 2 0 8 9>.

The transcriptions are uploaded as text files, represented in Unicode (utf-8) format. Several transcription files are allowed to be uploaded of the same cipher/key, and they should be uploaded as text files (.txt, docx, etc).

4.2 Cryptanalysis

The transcribed ciphertext can be analyzed by using various metrics. Attacking and eventually cracking ciphers might involve many different types of cryptanalysis. We implemented the ManuLab statistical analyzer (Antal and Zajac, 2018) for ciphers containing metrics for

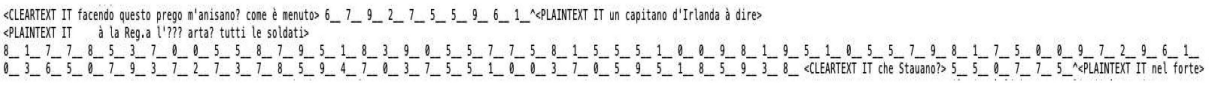

Figure 4: Transcription of the cipher image containing ciphertext, cleartext and plaintext in Figure 2.

index of coincidence, n-grams up to 8 characters, and entropy. The user can upload a ciphertext, run the appropriate metrics, and get analyzed data back. Then, the user can upload his or her own analysis to the database for the particular cipher, see Section 3.6. The analysis can be represented in terms of statistics, or a structured description of a study.

4.3 Mapping Ciphers to Keys

Given many ciphers and keys (being it original or generated), we developed a tool that maps a key/code/nomenclature and a ciphertext of the user's choice and creates a plaintext (i.e. decrypted text) on the basis of the provided key. The result is compared against 14 European historical language models, to see which language the plaintext matches best. The language models are provided by HistCorp, a collection of historical texts and language models for 14 European languages (Pettersson and Megyesi, 2018).

5 Searching in the Collection

The search function of the database allows for simple and advanced search depending on the user's need. Simple search, illustrated in Figure 5, allows for keyword searching, which looks for the occurrence of the search term in the record. The search is not sensitive to capitalization. The system matches the search term (i.e. the entered text) as a substring against all text fields with the exception of fields with checkbox and fields with numerical values. To search in these field types, the advanced search interface is used.

Advanced search allows the user to limit (or target) the search to each attribute or a combination of attributes defined by the metadata scheme with specific value(s). The Boolean operators AND, OR, and NOT are used in a graphical interface. The user first selects the attribute specified by the metadata. Most metadata fields use a text field for matching and the entered text is matched as a substring.

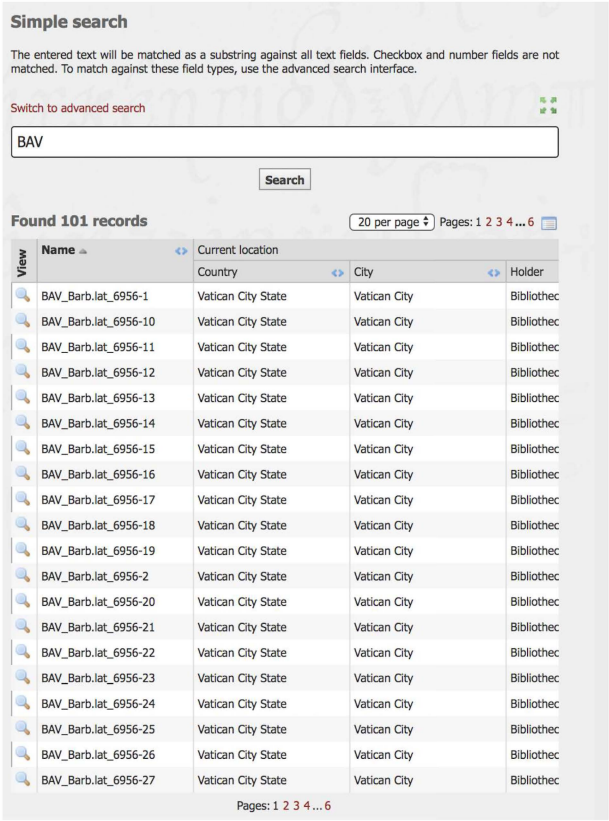

Figure 5: Simple search with matched records.

The NOT checkbox can be set to negate the expression. To the right of the selected field the user can delete the expression.

Search functions can be built upon each other for sequential search. The order of the operating functions are made explicit by grouping the expressions; composite expressions are shown in the same grey box appearing vertically, while consecutive operations, visualized horizontally, are executed in sequence. Figure 6 illustrates the advanced search function where we searched for all keys originated from either Germany or France from all times except from the years between 1700 and 1800.

The result of the search is shown either as a list of matched items line by line with metadata information shown in the columns for each item (see Figures 5 and 6), or as a

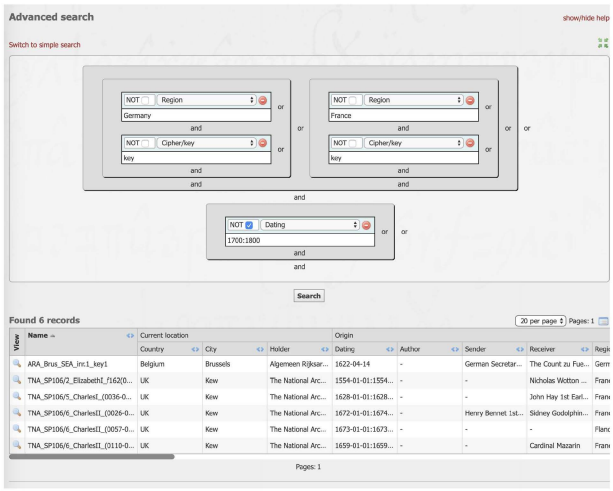

Figure 6: Advanced search.

"Google-style" result showing each record with its metadata listed item by item, see Figure 7. The magnifying glass shows all information about the particular record.

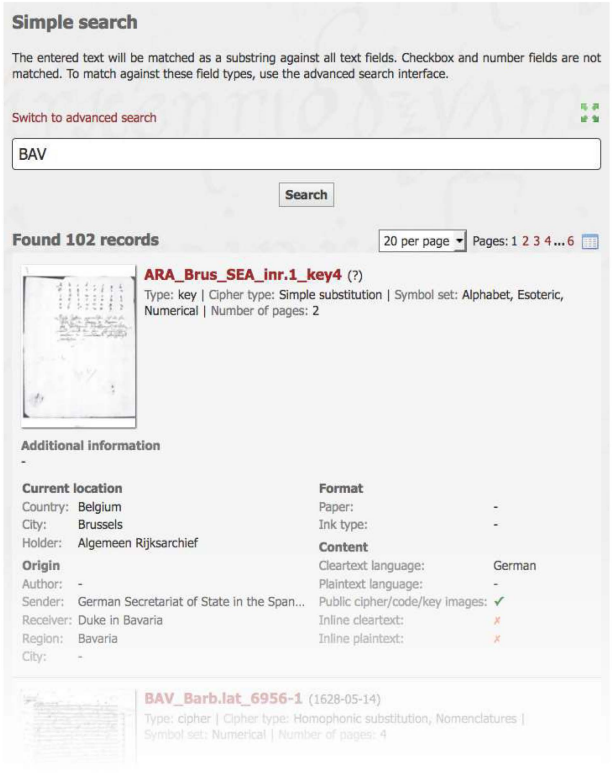

Figure 7: Search result.

Upon choosing a record from the list of matched documents, the record with all metadata is visualised, as illustrated in Figure 8 for the Copiale cipher with all metadata entered in the database.

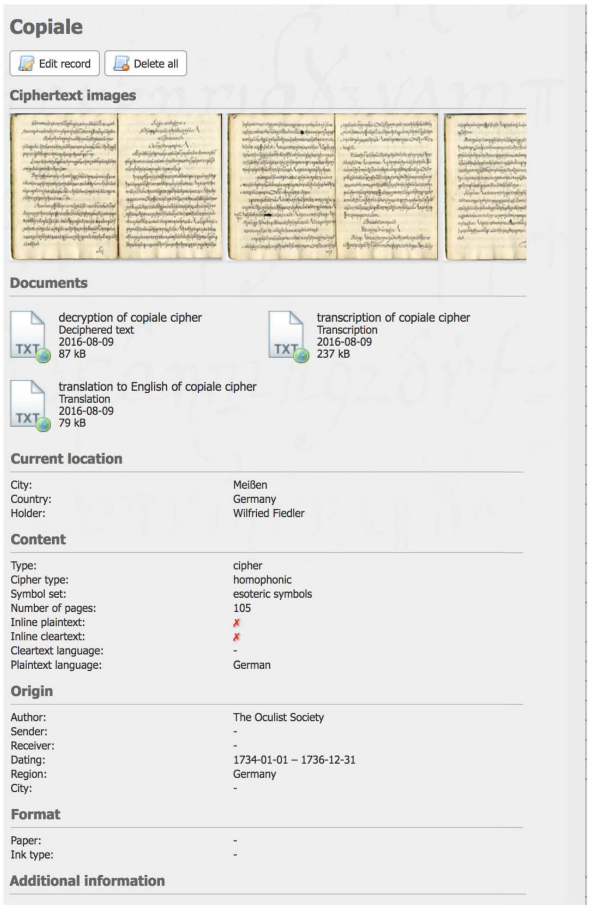

Figure 8: A cipher with metadata.

6 Adding New Data

One of the main goals of the DECODE database with its graphical user interaction is to allow registered users to create new records, ciphertexts or keys, by uploading an image of the encrypted document, and filling in metadata information about the manuscript. Mandatory fields are naming the manuscript, the current location (Country, City and Holder) of the manuscript, and the number of pages the manuscript consists of. All other metadata fields and related documents are optional.

For each uploaded document, being it an image, transcription, cryptanalysis, or publications, the user can choose to make it private, i.e. to not allow access to the file to other users.

7 Editing Existing Data

The user can edit information about existing records if she/he is the owner of the record. Uploaded documents, such as transcriptions, decrypted plaintexts, or translations of the

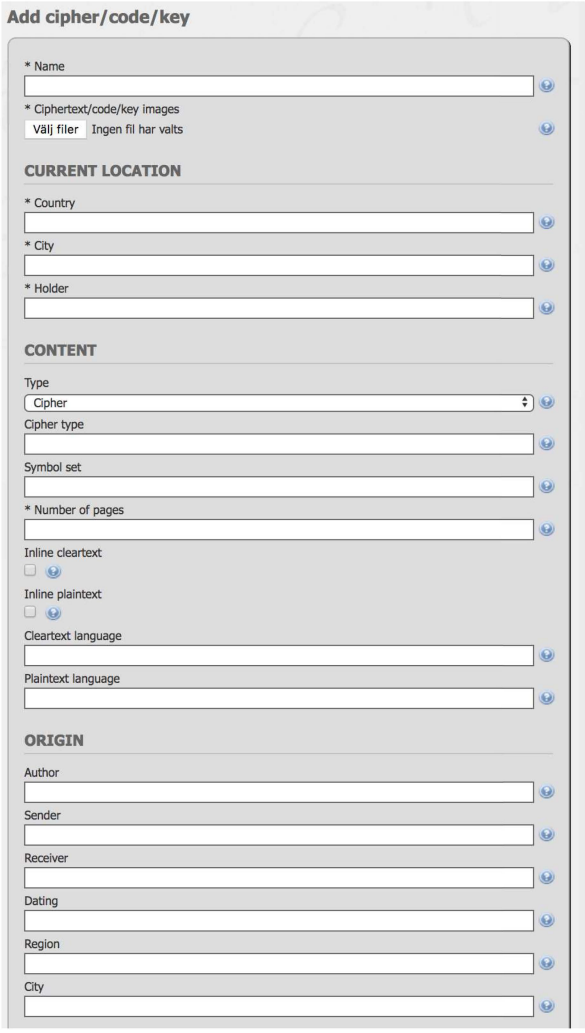

Figure 9: Entering new records.

plaintext can be deleted. New documents can be added and the metadata can be corrected. The owner can also choose to delete the whole record.

8 User Access and Roles

The DECODE database with a hosting web-service is publicly available with open access. In order to edit or add ciphers, codes or keys and related documents, registration is required of interested parties, typically specialists in historical cryptology. Personal data about the registered user includes information about the name of the user, affiliation, and scientific background (computational linguistics, computer science, cryptology, history, linguistics, literature, mathematics, politics, and other). The personal information that the user provides is handled according to GDPR. The personal information is stored and used to enable users of the site to see who has uploaded and modified certain content, and to simplify collaboration between users. The user may unsubscribe as a registered user upon request at any time without any explanation.

Registered users of the site may upload text, pictures and other information ("content") which is stored by DECODE, and is shared with other registered users and other users of the site, except in circumstances where registered users choose to make their content inaccessible to other users (i.e. private); this functionality is noted where available.

The user can remove the content that she/he uploaded (i.e. the owner of the record) at any time. Upon registration, the registered user agrees to not upload any content to which he/she does not hold the necessary rights. We do not claim any ownership rights of the content uploaded by the users. We do not use the content commercially. However, we share the content with third parties in order to perform image analysis and other types of analyses, except for the content that the user has chosen to make inaccessible. Should the content be in violation of applicable copyright law or other laws on intellectual property rights or be in any way abusive or illegal, we reserve the right to remove such content without any prior warning.

Should a user find any content on the site which is or may be in violation of applicable copyright law or other laws on intellectual property rights or in any way abusive or illegal, we ask the user to report this to us via email. Such information may be deleted by us at any time without any prior warning. The user has to tick a box that she/he agrees to the Terms and Conditions which are provided on the site upon registration, as described above.

9 Technical Description

The DECODE database web application is written in Python and runs as a WSGI application using Flask on the Apache web server via mod_wsgi. Secure session handling and logins are handled via the Flask extension Flask-Login. The codebase has two layers: The core logic, for which a test suite has been written to make sure that the system always remains in a consistent state, and the web part, which

calls into the application logic on behalf of the user.

All data itself (user data, record metadata) is stored in a PostgreSQL database, and is interacted with using Psycopg (a PostgreSQL-Python adapter). Image files are stored directly in the filesystem, although any access to these is done by first consulting the database. Search queries are served directly by the database.

Because Python does not handle dates before year 0 (ISO 8601), and because documents dated before that might be introduced into the database, the mxDateTime library from eGenix is used in place of the datetime module that the Python standard library provides.

Creation of thumbnails is done using Pillow, a fork of the Python Imaging Library. The 'pdfimages' utility from Poppler does the PDF image extraction, enabling users to directly upload PDF files of a scanned cipher instead of having to extract these themselves.

The server that the database web application is running on provides the SMTP connection, which is used to (via the Python standard library) send email to users, for example in order to reset one's password.

10 Tools

The database content is currently connected to CrypTool2 (CrypTool2, 2018) through an HTTP API, and connection to the Manulab system (Antal and Zajac, 2018) is underway.

To allow automatic processing of cipher images and transcription for decryption of the documents, we also develop historical language models extracted from authentic historical texts, and on-line tools for semi-automatic transcription, further develop cipher-key mapping, and the statistical analysis of ciphertexts. Currently, CrypTool2 is directly accessible for automatic decoding of ciphertexts where the user can test various decryption algorithms to decipher the encrypted elements.

11 Conclusion

We presented an on-line database aiming at the collection and systematic description of encrypted historical manuscripts. The database allows the user to search among ciphertexts and keys, and upload new encrypted historical sources with their metadata information and other relevant documents.

Future improvements include a mapping between ciphers and matching original and generated keys, and standardised forms for personal names, holding institutions and locations, like GND-number, VIAF or geonames-Number to be able to connect these in a systematic way for reliable search, as these entities are freely decided by the owner of the record today.

Most importantly, tools for automatic transcription, cryptanalysis and decryption are underway and will be connected to the database to allow self-help of the users.

Our hope is that many professionals interested in historical cryptology could make use of the data collected and enlarge and enrich the database with new ciphertexts, codebooks, and keys, or transcriptions, additional improved cryptanalysis, or in the best of worlds solution(s) to the undecrypted, still secret documents.

Acknowledgements

We would like to thank our colleagues Nicklas Bergman, Bengt Dahlqvist, and Per Starbäck for their help throughout the project, and all users and contributors who are willing to share these fascinating historical sources. This work was supported by the Swedish Research Council, grants E0067801 and 2018-06074.

References

Eugen Antal and Pavol Zajac. 2018. ManuLab System Demonstration. In *Proceedings of the 1st International Conference on Historical Cryptology, HistoCrypt 2018*, pages 125–128.

Archivio Segreto Vaticano ASV. 2016a. Asv16:segr.di.stato/portogallo/1a/16v@2016 archivio segreto vaticano. All rights reserved.

Archivio Segreto Vaticano ASV. 2016b. Asv16:segr.di.stato/portogallo/8/93r@2016 archivio segreto vaticano. All rights reserved.

CrypTool2. 2018. CrypTool2. last entry: 15/3/2019.

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.*