The BACK IN TIME Project

A User-Centred Platform for Encrypted Historical Documents

Cécile Pierrot
Inria, Nancy, France*
cecile.pierrot@inria.fr

Camille Desenclos
Université de Picardie
Inria, Nancy, France*
camille.desenclos@u-picardie.fr

Michaël Mera
Inria, Nancy, France*
michael.mera@inria.fr

Benjamin Kiessling
Inria, Paris, France
benjamin.kiessling@inria.fr

Gaspard Damoiseau-Malraux
Sorbonne Université, CNRS, LIP6
& Inria, Nancy, France*
gaspard.dama@lip6.fr

Hassen Aguili
Inria, Paris, France
hassen.aguili@inria.fr

Thibault Clérice
Inria, Paris, France
thibault.clerice@inria.fr

* is the short for:
Université de Lorraine
CNRS, Inria, LORIA
F- 54000 Nancy, France

Abstract

Since the Late Middle Ages, sensitive correspondence (political, diplomatic, military, etc.) has frequently been encrypted. Thousands of these documents now lie in archives and libraries, yet they remain largely beyond the reach of humanities researchers. The sheer volume and complexity of these materials render both manual cryptanalysis and manual analysis impractical, if not impossible. Although recent advances in historical cryptanalysis and artificial intelligence have enabled partial automation, existing tools remain ill-suited to the needs and practices of humanities researchers. The BACK IN TIME project was established in late 2024 to address these challenges. It seeks an alternative approach to automation that places the end user – namely, the humanities researcher – at the centre of the system's design. In this paper, we articulate our vision for a user-centred assistive platform intended to bridge the current gap: we highlight how the guiding principles identified by the project team can be used to combine computer vision, cryptanalysis, and historical expertise into a user-centred design that makes historical cryptanalysis both accessible and compatible with established scholarly practices. Initial implementations will focus on Western European scripts and languages, as well as on manually encrypted documents (up to the early 20th century), reflecting the team's expertise and easier access to these documents.

1 Introduction

While working with historical documents, especially historical correspondence, humanities researchers may encounter encrypted materials. These documents are unfortunately scattered and poorly identified within archives and libraries, and no systematic study of their volume and nature has yet been conducted. Existing works and projects suggest there may be thousands of encrypted documents, and correspondence of all types may have been encrypted, from diplomacy to intimacy (Láng, 2018; Héder and Megyesi, 2022; Tomokiyo, 2018). However, the presence of ciphertext does not necessarily prevent access to a document's content. Many encrypted documents have been preserved alongside the recipient's decryption, such as plaintext annotations in the margins, between lines, or on separate sheets.[1] In

[1] No systematic study has been conducted on decryption. According to our research, although it is unlikely that more than 1-2% of the encrypted documents never had a decryption (primarily intercepted letters with unsuccessful cryptanalysis), a significant portion (10-15%) has not been preserved

cases where plaintext is missing, the content may still be recovered (Desenclos and Lasry, 2025): it is possible to use the cipher key, provided the relevant key has both survived and been correctly identified. However, matching ciphertexts to their cipher keys requires significant time, extensive knowledge of archival repositories, and often depends on chance, as well-identified and preserved cipher keys are uncommon.[2] Even when keys are available or recovered, the subsequent manual decryption remains laborious and time-consuming,[3] particularly when handling larger or multiple corpora.

When part of the corpus contains both ciphertext and the corresponding plaintext, an alternative for humanities researchers is to reconstruct the cipher key themselves and then decrypt the remaining ciphertext. Although applied in several works (Stix, 1934 1936; Greenough, 1997; Monts de Savasse et al., 2004; Domnina, 2018; Pierrot et al., 2025; Benavent, 2025), this approach is also highly time-consuming, may require prior knowledge of cipher key structures, and poses a real problem when the corpus is large and/or when there are only a few short ciphertexts/plaintexts. Besides, it becomes outright impossible in the presence of a ciphertext alone (without a cipher key or plaintext) or of a ciphertext with some short plaintext but a complex or unknown encryption system.

Some tools (e.g., CTTS, CrypTool) can already assist manual cryptanalysis depending on the length and complexity of the ciphertext, while some works have investigated how to match cipher keys with ciphertexts (Pettersson and Megyesi, 2019) or ciphertexts with their plaintext (Bruton and Megyesi, 2025). In each case, the process remains a complex combination of automated techniques that are guided and supplemented by human intervention, and require expert knowledge. A first attempt to create a global pipeline was carried out by the DECRYPT project (Héder et al., 2024); however, this effort could not be completed. Despite their role as primary providers of encrypted documents and architects of historical research questions, humanities researchers still face a fragmented ecosystem of tools and methods that, save for a few highly specialised experts with extensive cross-disciplinary knowledge, they cannot use due to a lack of technical computer skills and unfamiliarity with historical ciphers.

To enable humanities researchers to access the content of encrypted documents and better understand encryption practices and systems, an alternative approach is required. Since its inception in late 2024, the BACK IN TIME project has investigated the benefits of designing an automated yet user-centred cryptanalysis process. Alongside the development of tools for automated decryption, the core of the BACK IN TIME project is the realisation of a user-centred pipeline that enables the end-user – the humanities scholar – to process encrypted documents autonomously. In this paper, we reassert the importance of this goal and outline our findings about the core principles that will ultimately enable the development of such a pipeline.

2 Goals

The BACK IN TIME project aims to address both the significant technical challenges presented by automated historical document decryption (technical tools) and the prerequisites for autonomous usability by humanities researchers (user-centred pipeline) simultaneously. The latter aspect is the main methodological basis of the project, driven by the fact that such collaboration is a fundamental precondition for the very creation of valid, usable and used research tools. Humanities researchers are indeed uniquely positioned to further the development of automated decryption by providing relevant ciphertexts, high-quality, accurate and structured historical metadata (such as date, language, origin, and historical context), and valuable feedback both on the transcription (palaeography) and on the reconstructed plaintext (historical and linguistic accuracy). To encourage this complementary expertise, the project values developing a platform that serves as a collaborative hub for end-users and the diverse communities involved in the automation process.

2.1 The Platform's Design Principles

Along with studying existing software and developing new tools for automated decryption, we identified the following core principles for a plat-

alongside or close to its decryption (plaintext) and requires cryptanalysis.

[2] Many cipher keys are actually matched to their ciphertexts thanks to cryptanalysis. (Kopal and Waldispühl, 2021; Desenclos and Lasry, 2024)

[3] According to the experiments we conducted, a researcher unfamiliar with historical ciphers will need, on average, 1 hour and 45 minutes to decrypt one fully encrypted page (approx. 30 lines)

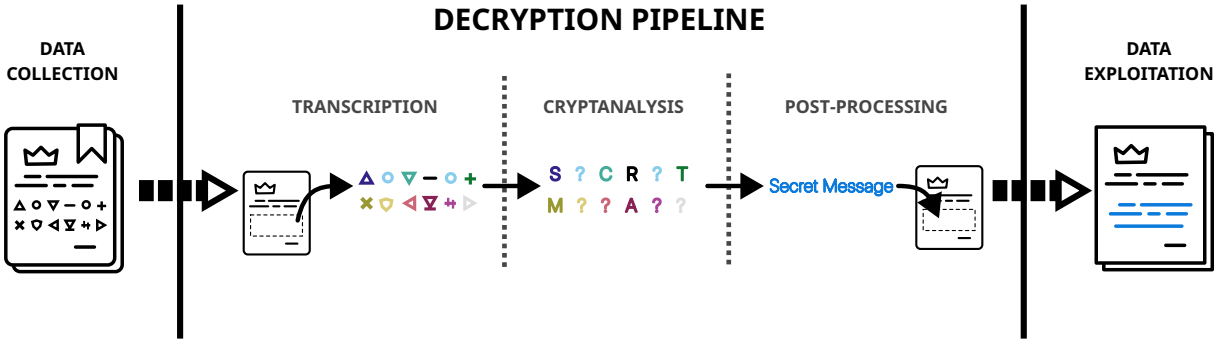

Figure 1: The standard pipeline for the processing of encrypted documents.

form centred on end users.

End-to-End: A significant barrier to the widespread adoption of computational methods for processing encrypted documents is the fragmentation of available tools, each designed to address only isolated stages of the processing pipeline. Furthermore, some specific components, such as ciphertext character recognition, remain important research challenges and are not yet fully automated tasks (Souibgui et al., 2023; Fornés et al., 2024). While this disjointed landscape may be navigable for experts already well-acquainted with historical cryptography and its necessary processing steps, it presents an almost insurmountable obstacle for novice and occasional users.

To address this fragmentation, we suggest integrating our tools and providing a single user-facing interface that spans the entire pipeline. This end-to-end design ensures that users can concentrate on their primary domain of expertise, specifically the collection of historical documents and the interpretation of reconstructed plaintext, which are situated at the input and output sides of the pipeline (see Figure 1). The complexity of the intermediate steps should be abstracted away to create a unified workflow designed specifically for humanities researchers (without a technical background).

User-centred: While a single interface already enhances consistency, it must also provide a clear and intuitive workflow for a wide target audience without prior knowledge in historical cryptography. In practice, this workflow must transpose the specialised terminology and abstract away algorithmic details. An informal survey conducted among humanities researchers (history, literature) has shown that successful use of such a platform and its integration within the humanities ecosystem requires tools that are grounded in both vocabulary and methodology in existing scholarly norms, tools that provide accessible and introspectable outputs at every stage of the transcription and cryptanalysis process. Within the BACK IN TIME platform, cryptanalysis algorithms and computer vision AI models, for example, will be handled internally and not exposed to the end-user. Of course, this is not to say that the user will be passive during the process, but rather that the feedback will be requested in terms that match the user's background, e.g. *Are the following characters correctly identified?* or *Is the following text correct in the language of the message?* among others.

User-in-the-loop: Designing a fully automated process for the decryption of historical documents poses significant challenges for both computer vision and cryptanalysis. At the time of writing, no such comprehensive system exists. To address these challenges, the user will need to be fully involved, assisting and guiding the process. Rather than waiting until the end of each pipeline step to query the user's (dis)approval, it is critical to design a workflow that actively and regularly collects feedback throughout the entire processing pipeline. It allows for a more fine-grained and dynamic parameterisation of the pipeline, as highlighted in subsection 3.3. However, while designing such a workflow requires abstracting away the technical intricacies of computer vision and cryptanalysis, it is equally important to let users understand the role of each functional block. This enables them to better identify error sources and apply their expertise (linguistic, historical) to redirect the process in the right direction when needed.

Modular: Given the extreme heterogeneity of encrypted historical documents, the assumption that a single algorithm or AI model could universally address such diversity is unrealistic. As such, the workflow is built upon the foundational principle of modularity. Rejecting a monolithic architecture with fixed implementations and general models, the platform is designed to integrate a diverse array of interchangeable modules. Each module targets a specific situation, such as a particular encryption system, secondary security devices (nulls, nomenclature, etc.), ciphertext character sets, plaintext script and language, allowing the workflow to be dynamically configured to suit the specific requirements of each document or corpus.

Resilient: Despite its recent growth, there are still gaps in the current research that will inevitably lead to failures on the platform. Due to a lack of metadata or knowledge, some types or parts of encryption systems may not be correctly identified by the automated process or, due to the rarity or complexity of some encryption systems, no generic cryptanalysis algorithm would work on them. Some encrypted documents present interlinear plaintext that the computer vision algorithms might not correctly recognise during segmentation, or ciphertext characters that the algorithms might not correctly identify. As such, there will be documents that cannot be successfully handled by any of the modules provided on the BACK IN TIME platform. Rather than trying to solve everything within our funded time, we only address the most recurring challenges (ciphertext character sets and encryption systems) and design a platform to gracefully handle failure cases. In case of failure, the type of problem encountered should be automatically reported to the project team, and the particular use case will be analysed with the user to identify its specificities. In the longer term, a dedicated expert module that solves the problem could be developed to increase the platform's coverage.

Open Source, Free, Transparent: We strongly believe that only an open-source and freely accessible ecosystem can foster durable, sustained progress in research. Open science is not merely a matter of transparency, but a prerequisite for reproducibility, collective validation, and long-term impact, especially in interdisciplinary fields where methods and data must be scrutinised, adapted, and extended by diverse communities. By releasing the source code developed throughout this project, we aim to facilitate reuse, critical assessment, and incremental improvement across all kinds of expertise (history, computer vision, cryptanalysis, etc.). Beyond providing open access to the code, we plan to host a free instance of the platform in order to lower the barrier to entry for users who lack the technical infrastructure or expertise to deploy such tools themselves.

Responsible Data Management: Beyond mere compliance with the European General Data Protection Regulation, we need to establish a rigorous data management policy designed to ensure users maintain absolute stewardship over their materials. The platform should operate on a strict non-retention basis by default: all uploaded documents remain the exclusive property of the user and are neither stored, redistributed, nor utilised by the project without authorisation.

Conversely, researchers may choose to contribute their research artefacts (transcriptions, cryptanalysis, recovered cipher keys) to the community, but only through a conscious, explicit opt-in mechanism. This commitment to user autonomy is critical for building trust, particularly given that archival materials may be subject to restrictive reproduction and dissemination rights. Through these principles, the project ensures that the user remains the sole arbiter of how their data is shared and valued.

2.2 Community-Driven Development and Open Challenges:

User-centred or not, the pipeline needs to rely on technical expertise from diverse technical fields. But, if the BACK IN TIME project involves the development of specialised software tools, these tools actually aim to simplify the creation of expert modules, allowing the specific problem-solving strategies of computer scientists and cryptographers to be integrated directly into the historian's workspace. By lowering the barrier to technical contributions, we aim to transform the platform into a nexus where these distinct forms of expertise can coalesce into a shared methodology. We posit that the complexity of historical, especially automated, cryptanalysis cannot be resolved solely through the research methodology of a single domain. Instead, it requires a synergistic ap-

proach that converges the distinct inputs of cryptanalysts, computer vision experts, and historians into a unified operational framework.

To encourage and support this collective dynamic, we drew inspiration from existing initiatives that use challenges to leverage external expertise to advance some research questions and challenges in the specific field of historical cryptography.[4] These challenges have proved their value. In the BACK IN TIME project, we intend to propose different kinds of challenges, such as code-breaking tasks on synthetic data, to calibrate cryptanalysis algorithms more precisely. However, our main focus will be on encouraging humanities researchers to submit their own challenges (e.g. an encrypted document that they cannot decrypt) to the platform. We hope to achieve several goals in this way: bringing humanities researchers to the platform (as active contributors and not as passive users) even before the release of the full automation pipeline, enabling the continuous refinement of transcription and cryptanalysis algorithms (multiple examples and valuable feedback), and addressing specific cryptographic challenges (unknown or rare encryption systems, rare ciphertext characters, etc.). Moreover, these challenges will enable researchers to identify the most effective solutions for specific tasks within our automation decryption pipeline.

By fostering collaboration around well-defined problems, the challenges will help verify that the pipeline remains aligned with the state of the art and users' needs while benefiting from the collective intelligence of its diverse members. Moreover, by prioritising the construction of open, community-driven infrastructures, we ensure that methodological advances are not lost when the project ends, but remain accessible and verifiable. This shared ecosystem is essential to the field's long-term sustainability.

2.3 Studying the Gap in Current Research

As outlined in subsection 2.1, the platform must be designed with resilience and modularity in mind, to avoid enforcing a *one-size-fits-all* model and instead to make processing errors explicitly visible and identifiable. In such a framework, a failure is indeed an informative signal, rather than a technical glitch. In its current state, history of cryptology can only provide an incomplete taxonomy of encryption practices and systems. Therefore, failures of proven algorithms are more generally hints of limitations of the taxonomy itself, rather than of the effectiveness of said algorithms.

Failures indicate that the material deviates from expected norms, whether in its visual structure (rare or badly formed ciphertext characters), cryptographic logic, or language (encryption errors, dialect, etc.). Hence, this signal serves as a heuristic for discovery, highlighting the necessity for a dual research effort: the development of more robust computational methods capable of handling outliers, and deeper historical research to better describe and categorise the vast diversity of extant sources.

This diagnostic approach is particularly critical given that current computational methods are rarely systematically evaluated against the full diversity of historical material. This lack of comprehensive benchmarking is understandable, stemming from the scarcity of representative datasets and the genuine heterogeneity of real-world encryption practices. By systematically capturing these failure cases, the platform enables a finer-grained analysis of algorithmic boundaries. It allows researchers to relate performance not just to abstract metrics but to concrete document properties, such as ciphertext character sets or layout characteristics, thereby transforming implicit algorithmic limitations into explicit and actionable research questions.

3 User-centred Pipeline

The platform we envision aims to provide a strict separation of concerns between the interactions exposed to the end-user and the decisions involving domain-specific automation tools and algorithms. To guarantee this separation, we structure the platform using a layered architecture, as shown in Figure 2.

3.1 User Interface

At the very top of the layered architecture, the user interface displays a unified workflow decoupled from the underlying algorithms and expert modules involved in the automated decryption process. This will let the user focus on the various

[4] Many websites propose cryptanalysis challenges (e.g., `https://mysterytwister.org/`) or unsolved cryptograms (e.g., `https://crypto.hcportal.eu/dashboard/cryptograms`). More recently, challenges dedicated to Handwriting Recognition of Historical Ciphers have also arisen (Fornés et al., 2024).

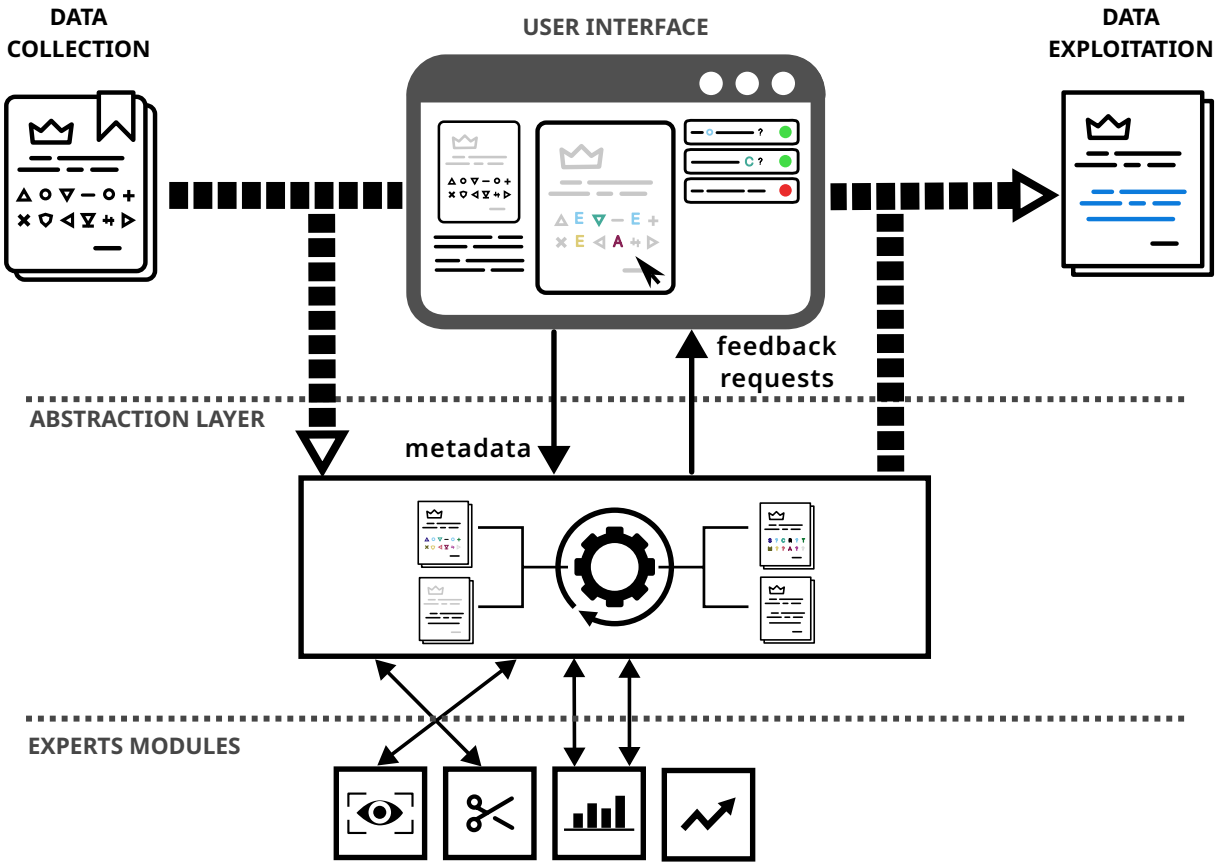

Figure 2: The layered architecture for the user-centred platform.

outputs rather than the mechanisms that drive the transcription and cryptanalysis. For instance, in terms of the final result, it should not matter to the user which type of AI model, which cryptanalysis algorithm or combination of algorithms has been used to perform the decryption. On the contrary, it should matter to the user to get classification both of the ciphertext characters and of the encryption system, as well as to know whether the plaintext has been reconstructed by pure cryptanalysis, by matching an already stored cipher key or by matching existing plaintext.

More specifically, the user interface integrates the entire pipeline and presents it to the user as an iterative workflow: Starting from a digitised version of the encrypted document (e.g. personal photo, online digitised documents), the document's text is progressively extracted (segmentation then transcription), cryptanalysed and refined into a final usable plaintext alongside the reconstructed cipher key. This iterative workflow is sustained by the user, who responds to feedback requests from the underlying layers, allowing the platform to gradually refine the result while collecting additional metadata. At each stage and iteration, the user will also have the option to obtain specific, workable outputs to independently analyse the ciphertext characters (classification, frequency statistics, visualisation of usage) as well as the encryption system (reconstructed cipher key and parameters).

3.2 Expert Modules

The collection of expert modules resides at the bottom of the layered architecture, serving as the platform's foundational engine. The primary role of an expert module is to encapsulate specific technical logic required to solve a standalone task, such as performing cryptanalysis on a specific cipher type or segmenting lines of text from a noisy image. These modules are typically implemented as wrappers around existing methods developed by domain experts, such as cryptologists or computer vision specialists, integrating diverse external software into a unified ecosystem.

To ensure seamless integration, each module must adhere to a standardised interface. However, because no single algorithm can handle the vast diversity of historical documents and encryption methods, strict scope declarations are mandatory.

These declarations act as machine-readable contracts that specify the exact prerequisites for the module's operation. For instance, a cryptanalysis module must explicitly state which encryption systems it supports (e.g., homophonic ciphers) and the valid ranges for key parameters. Similarly, an image processing module designed for handwriting recognition might restrict its scope to specific script types, such as Roman numerals or abstract symbols, while rejecting printed text.

This modularisation presents a significant opportunity for standardisation. By forcing domain experts to formalise the capabilities and limitations of each tool via scope declarations, the platform creates a definitive list of criteria for algorithm evaluation. Furthermore, as these modules are deployed against real-world data, their scope definitions can be iteratively refined, narrowing the gap between theoretical capability and practical application.

3.3 Abstraction Layer

The abstraction layer functions as intelligent middleware connecting the high-level user interface (subsection 3.1) with the low-level expert modules (subsection 3.2). Its primary responsibility is orchestration: it must interpret the current state of the document processing, identify blocking issues (such as unknown or uncommon encryption systems, poorly recognised or overlooked ciphertext characters, unintelligible reconstructed plaintext), and dynamically select the most appropriate expert modules to resolve them.

This selection process relies heavily on the scope declarations provided by the underlying modules. The abstraction layer matches the metadata and features of the encrypted document against the available modules to determine eligibility. For a simple encrypted document (easily identifiable encryption system, one or a few security devices), this may involve only a linear sequence of calls: segmentation, followed by recognition, followed by cryptanalysis, etc. However, for a complex encrypted document (with a partially or non-identified encryption system or multiple security devices), the abstraction layer employs sophisticated resolution strategies. It may act as a competitive orchestrator, invoking multiple cryptanalysis modules in parallel to compare their outputs, or it may apply a fallback strategy, trying a computationally expensive algorithm only after a faster heuristic has failed.

Crucially, the abstraction layer is also responsible for data synthesis. It must harmonise the raw, often technical outputs from various modules into a coherent structure that the User Interface can render. By presenting synthesised and readable results to the user for validation, the abstraction layer creates a feedback loop: if the user rejects a result, the layer can interpret this as a constraint update, triggering a new search for alternative modules or requesting specific input to unlock advanced processing paths.

4 Related Work

The BACK IN TIME project is neither the first nor the only one dedicated to the processing of encrypted historical documents. Key initiatives paved the way for identifying major methodological challenges and research questions and facilitated the development of a few pioneer tools. While the insights and outcomes of these efforts have deeply informed our conceptual framework, BACK IN TIME does not intend to duplicate or repackage existing tools. Instead, to ensure the end-user remains the focal point of the global pipeline, the project explores an alternative design strategy. This entails the development of novel tools from the ground up, while strategically incorporating and adapting established algorithms, e.g., eScriptorium's layout analysis functionality (Kiessling, 2020), and capitalising on recent technical advancements in machine learning, newly combined cryptanalysis algorithms, as well as the latest work on the history of cryptology to overcome the limitations identified with the current research landscape.

The following overview examines initiatives such as the DECRYPT (Megyesi et al., 2020) and DESCRYPT (Megyesi et al., 2025) projects as well as existing tools and environments, including CrypTool2[5] for cryptographic experimentation, Cryptool transcriber and Solver (CTTS)[6] for manual transcription and cryptanalysis, and eScriptorium[7] (Kiessling et al., 2019) and Transkribus[8] for large-scale transcription. By analysing these contributions, we identify their respective strengths and limitations, thereby clarifying how our ap-

[5] https://www.cryptool.org/en/ct2/
[6] https://www.cryptool.org/en/ctts/
[7] https://escriptorium.inria.fr/
[8] https://www.transkribus.org/fr

proach builds upon and diverges from the existing landscape.

Historical Cryptanalysis Projects. Since the mid-2010s, a few initiatives have emerged to advance the computational analysis of encrypted historical documents, most notably the DECRYPT project (Megyesi et al., 2020). These efforts have significantly advanced the study of history of cryptology and structured the field of historical cryptography by developing and integrating a wide array of tools and resources. As does BACK IN TIME, the DECRYPT project addresses the primary stages of processing encrypted documents, with research activities ranging from data collection (via the DECODE database) to cryptanalysis. The final phase of DECRYPT (Héder et al., 2024), as well as the workflow of the new DESCRYPT program[9], envisions linking their various tools and resources into a unified pipeline. However, this integrated workflow has not yet been fully realised (Héder et al., 2024). Furthermore, the proposed architecture assumes a more linear pipeline, applying modules one after another, which contrasts with the iterative and exploratory process often required by humanities researchers. The BACK IN TIME team pursues a very different strategy. By prioritising a unified platform from the outset, we aim to demonstrate that placing a user-centred, iterative workflow at the centre of the architecture yields genuine benefits for developing new tools and integrating them into research practices.

Further, even though DECRYPT has undeniably demonstrated the feasibility of applying computational methods to historical ciphers, particularly by developing algorithmic approaches for cryptanalysis and establishing experimental frameworks for transcription, its software ecosystem remains primarily designed for expert use. Key tools provide robust environments for cryptanalysis but are distributed as standalone executables requiring specific technical environments. For instance, CrypTool Transcriber & Solver (CTTS) is a Java-based desktop application that requires command-line interaction and manual configuration, with documentation hosted on GitLab, a platform unfamiliar to most humanities scholars. Similarly, tools for semi-automatic transcription, such as Transcripttool,[10] impose comparable technical barriers. This overhead, combined with interfaces that presuppose prior knowledge of cryptography, e.g., requiring the user to select specific cipher parameters before analysis, renders these tools, on their own, unsuitable for humanities researchers whose primary objective is recovering document content rather than acquiring expertise in historical cryptography. However, to ensure compliance with all the platform's design principles (subsection 2.1) and to integrate the latest technical and methodological advances, new tools (transcription and cryptanalysis) will be designed for the BACK IN TIME pipeline.

Community Resources and Datasets. Part of the DECRYPT ecosystem, the DECODE database (Héder and Megyesi, 2022) is a unique example of a collaborative and searchable database of encrypted documents and cipher keys. By making such materials available, it has enabled a more systematic analysis of encryption practices and systems across time (Megyesi et al., 2021). Such work will not be integrated within BACK IN TIME: the aims are different – the humanities researchers are directly providing encrypted documents they need to decrypt and/or analyse – and DECODE is already widely acknowledged as the main repository for encrypted documents and cipher keys. Similarly, large-scale historical corpora such as HistCORP[11] (Pettersson and Megyesi, 2018) play a crucial role by offering curated resources that support comparative and contextual analyses in historical research. Although neither DECODE nor HistCORP will be integrated into the BACK IN TIME architecture, they provide a foundational data layer that complements the data our team already collected, resulting in a diverse corpus for training and evaluating our transcription and cryptanalysis tools.

Transcription Platforms for Historical (non-encrypted) Documents. Platforms such as eScriptorium (Kiessling et al., 2019) and Transkribus have reshaped the field of digital humanities, unlocking large-scale transcription of handwritten manuscripts. Driven by advances in deep learning, they have significantly lowered the barrier to producing accurate transcriptions from digitised

[9] https://descrypt.org/ The DESCRYPT project reuses and extends the tools and expertise of DECRYPT to cover rare scripts and low-resource texts.

[10] https://github.com/decrypt-project/transcripttool

[11] https://www2.lingfil.uu.se/person/pettersson/histcorp/

handwritten texts.

However, the specific nature of encrypted documents introduces fundamental challenges beyond the scope of standard transcription platforms. First, modern handwritten text recognition architectures rely heavily on linguistic modelling to compensate for visual ambiguity[12] an approach that fails when applied to ciphertexts designed specifically to maximise entropy and eliminate predictable lexical patterns. Second, ciphertext characters can differ significantly from cleartext scripts in their visual composition, necessitating the creation of substantial new ground-truth datasets to train models on non-standard character sets. Finally, whereas standard models depend on fixed character inventories and closed classification layers, symbolic ciphertext characters often operate as open or semi-open sets, which some have approached by returning to basics and character classification (Siglidis et al., 2024).

Consequently, while we cannot use transcription platforms out of the box, BACK IN TIME adopts a hybrid strategy. It integrates the eScriptorium infrastructure specifically for its state-of-the-art layout analysis and line segmentation capabilities (the first stage of the transcription pipeline). However, it replaces the subsequent text recognition block with its own specialised modules designed to handle the visual complexity and linguistic entropy of encrypted script.

5 Conclusion

The BACK IN TIME project certainly addresses already-identified issues, such as the transcription of encrypted documents, their cryptanalysis, and the development of an automated decryption pipeline. Some initiatives already dedicate resources to some of these issues, but offer a fragmented, hardly usable ecosystem for humanities researchers. However, beyond designing new tools, it plans to explore a different, user-centred approach by revising the classical pipeline and rebuilding each stage to fit the needs of the end user: humanities scholars. We hope this will reinforce automation efforts while being progressively integrated into the toolbox of humanities researchers who can benefit from and could more often contribute to historical cryptography. While this project does not intend to solve every kind of cryptogram and will first focus on the most commonly used ciphertext character sets and encryption systems, it aims to make continuous improvements using a federated approach that brings together all researchers working with encrypted historical documents. The ultimate goal is to enable humanities researchers to engage directly with cryptographic sources at scale, thereby transforming encrypted documents from marginal curiosities into fully exploitable historical evidence.

[12] TrOCR (Li et al., 2023) includes a language transformer after a visual one, PyLaia (Tarride et al., 2024), while older, provides connector for small language models KenLM (Heafield, 2011).

Aknowledgement

This work received government funding managed by the French National Research Agency (ANR) under France 2030, reference number ANR-24-RRII-0002, named *Back In Time* project and operated by the Inria Quadrant Programme (PIQ).

References

Júlia Benavent. 2025. La escritura secreta en Europa en el siglo XVI. *Estudios románicos*, 34:11–189.

Micaella Bruton and Beáta Megyesi. 2025. From Statistics to Neural Networks: Enhancing Ciphertext-Plaintext Alignment in Historical Substitution Ciphers for Automatic Key Extraction. In *Proceedings of the 8th International Conference on Historical Cryptology, HistoCrypt 2025*, pages 18–26. Linköping University Electronic Press.

Camille Desenclos and George Lasry. 2024. An early French digit cipher: deciphering a letter from the King of France to the Duke of Nevers (1592). In *Proceedings of the 7th International Conference on Historical Cryptology (HistoCrypt 2024)*, pages 46–56. Tartu University Library.

Camille Desenclos and George Lasry. 2025. Cryptanalytic and historical challenges with unidentified encrypted documents from the early modern era. In *Proceedings of the 8th International Conference on Historical Cryptology, HistoCrypt 2025*, pages 41–51. Linköping University Electronic Press.

Ekaterina Domnina. 2018. Nicodemo Tranchedini's Diplomatic Cipher: New Evidence. In *Proceedings of the 1th International Conference on Historical Cryptology (HistoCrypt 2018)*, pages 3–7. Tartu University Library.

Alicia Fornés, Jialuo Chen, Pau Torras, Carles Badal, Beáta Megyesi, Michelle Waldispühl, Nils Kopal, and George Lasry. 2024. ICDAR 2024 Competition on Handwriting Recognition of Historical Ciphers. In *Document Analysis and Recognition - ICDAR 2024*, pages 332–344. Springer Nature Switzerland.

H. Paul Greenough. 1997. Cryptanalysis of the Swedish HC-9. A Known-Plaintext Approach. *Cryptologia*, 21(4):353–367.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

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)*, pages 111–114. Tartu University Library.

Mihály Héder, Alicia Fornés, Nils Kopal, Ferenc Szigeti, and Beáta Megyesi. 2024. Supporting Historical Cryptology: The Decrypt Pipeline. In *Proceedings of the 7th International Conference on Historical Cryptology (HistoCrypt 2024)*. Tartu University Library.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stokl Ben Ezra. 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19. IEEE Computer Society.

Benjamin Kiessling. 2020. A modular region and text line layout analysis system. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 313–318. IEEE.

Nils Kopal and Michelle Waldispühl. 2021. Deciphering three diplomatic letters sent by Maximilian II in 1575. *Cryptologia*, 46(2):103–127.

Benedek Láng. 2018. *Real Life Cryptology: Ciphers and Secrets in Early Modern Hungary*. Amsterdam University Press.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13094–13102.

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*, 44(6):545–559.

Béata Megyesi, Crina Tudor, Benedek Láng, and Anna Lehofer. 2021. Key Design in the Early Modern Era in Europe. In *Proceedings of the 4th International Conference on Historical Cryptology (HistoCrypt 2021)*, pages 121–130. Tartu University Library.

Beáta Megyesi, Alica Fornés, Mihaly Héder, Raphaela Heil, Nils Kopal, Benedek Láng, Rune Rattenborg, and Michelle Waldispühl. 2025. Decipherment of Historical Manuscripts with Unknown or Rare Writings: The DESCRYPT Project. In *Proceedings of the 8th International Conference on Historical Cryptology, HistoCrypt 2025*, pages 101–105. Linköping University Electronic Press.

Jacques de Monts de Savasse, Yves Soulingeas, and Stéphane Gal. 2004. *L'Europe d'Henri IV. La correspondance diplomatique du Secrétaire d'Etat Louis de Revol (1558-1593)*. Presses Universitaires de Grenoble.

Eva Pettersson and Beáta Megyesi. 2018. The HistCorp Collection of Historical Corpora and Resources. *Digital Humanities in the Nordic and Baltic Countries Publications (DHN2018)*, 1:306–320.

Eva Pettersson and Beáta Megyesi. 2019. Matching Keys and Encrypted Manuscripts. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 253–261. Tartu University Library.

Cécile Pierrot, Gaspard Damoiseau-Malraux, Paul Mekhail, Olivier Chaline, and Ludovic Perret. 2025. A Caribbean Directory-based Encryption during the American War of Independence. In *Proceedings of the 8th International Conference on Historical Cryptology, HistoCrypt 2025*, pages 122–132. Linköping University Electronic Press.

Ioannis Siglidis, Nicolas Gonthier, Julien Gaubil, Tom Monnier, and Mathieu Aubry. 2024. The learnable typewriter: a generative approach to text analysis. In *International Conference on Document Analysis and Recognition*, pages 297–314. Springer, ICDAR.

Mohamed Ali Souibgui, Pau Torras, Jialuo Chen, and Alicia Forné s. 2023. An Evaluation of Handwritten Text Recognition Methods for Historical Ciphered Manuscripts. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, page 7–12. Association for Computing Machinery.

Franz Stix. 1934–1936. Die Geheimschriftenschlüssel der Kabinetskanzlei des Kaiser. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen Philologisch-Historische Klasse*, 1–2:207–226 and 61–70.

Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving Automatic Text Recognition with Language Models in the PyLaia Open-Source Library. In *International Conference on Document Analysis and Recognition (ICDAR 2024)*, pages 387–404. Springer Nature Switzerland.

Satoshi Tomokiyo. 2018. Cryptiana: Articles on Historical Cryptography.