OTRUCH (Outil de TRanscription Universelle de Corpus Historiques)

Otruche2025!

Project title
Scientific **T**ext **A**nalysis and **R**ecognition
Project acronym
STAR
Project summary (1200 car)

This project tackles long-standing methodological issues in the history of science by integrating advanced computer vision technologies and digital humanities with historical expertise. Despite progress in the field, the history of science remains constrained by 19th- and 20th-century frameworks, including Eurocentric periodizations, disciplinary boundaries, and fragmented approaches to textual sources. To address this, the project relies on a large-scale corpus of diverse scientific texts—manuscripts and printed works—from Asia, Africa, and Europe (8th–18th centuries). It emphasizes the discursive, word-based elements of texts that link and structure non-discursive components like diagrams and tables. While current OCR and HTR technologies are too specialized and fragmented, posing challenges for diverse textual traditions, the project will create a unified data model and automatic transcription interface, along with a universal approach to selecting the best OCR/HTR models. By merging historical and computer vision expertise, it aims to challenge disciplinary boundaries, advance digital humanities, and open new pathways for global comparative research in the history of science.

Keywords (3-5)
HTR, Scientific sources, digital platform

Objectives and description

History of science has, over the past several decades, drawn on an increasingly diverse array of methodological inspirations — including history, philosophy, sociology, and anthropology. It has also undergone various methodological "turns," such as the “practical turn,” the “material turn,” and the “visual turn.” Through these developments, the field has often aligned itself with, and made significant contributions to, broader historiographical currents such as environmental history and connected history. Nevertheless the history of science remains constrained by significant bottlenecks. These bottlenecks are rooted in choices and approaches established during the late 19th and early 20th centuries, when the discipline first began to take shape. Among these enduring constraints are the tendency to study historical sciences through the lens of contemporary disciplinary boundaries, the framing of scientific practices as expressions of discrete geographical or civilizational units, and the reliance on Eurocentric periodization schemes. Such limitations often lead to oversimplified narratives of "precursors" and "scientific heroes" while reinforcing rigid boundaries within the discipline itself.

Our proposal, therefore, is to address these bottlenecks directly—but not merely in an abstract or theoretical manner. We hypothesize that these constraints are not only historiographical in nature but are also deeply embedded in the discipline's fundamental working practices, particularly in its engagement with primary sources and access to historical texts. The aforementioned bottlenecks are reflected in the fragmented approaches to textual sources in the history of science. For example, philological practices and critical standards often differ significantly across various domains of sources, fundamentally obstructing the possibility of constructing robust historical analyses that rely on sources spanning these diverse domains. Building on the most recent advancements in computer vision and digital humanities, and drawing on a unique combination of historical expertise, our intervention will focus on developing new digital tools. These tools will enable historians of science to work simultaneously with a broad range of manuscripts and printed sources, encompassing the widest possible chronological and geographical scope. In doing so, we aim to challenge the foundations of the identified bottlenecks and open up entirely new research pathways for the discipline.

While fully embedded within the writing cultures of their respective historical contexts—whether manuscript, print, or digital—scientific sources often exhibit distinctive features that set them apart from other written artefacts. A central characteristic in this regard is the pronounced reliance of scientific texts on non-discursive elements, such as illustrations, diagrams, maps, and tables. Moreover, these texts frequently develop not only specialized terminologies but also discipline-specific forms of writing, such as mathematical formulas. These features, in turn, influence the use of more conventional discursive, word-based elements within scientific sources. Notably, these discursive elements frequently act as a kind of "connective tissue" within the source, structuring and infusing meaning into the work as a whole. In previous and ongoing projects, I have investigated non-discursive elements in scientific texts—such as tables (*DISHAS*) and diagrams (*EIDA*)—particularly within the context of astronomical sources. With this proposal, however, I aim to initiate a new approach for the exploration of discursive elements of scientific sources. To appreciate this approach's novelty, we must briefly assess the digital humanities context and current OCR/HTR technologies, which are key to our objectives.

In the field of digital humanities, significant research efforts and remarkable progress have been made in the treatment of discursive, word-based texts. The Text Encoding Initiative (TEI) is increasingly becoming a standard, serving both as a foundation for digital publication and as a format for many natural language processing (NLP) applications. Similarly, SegmentOnto is establishing itself as a standard for page segmentation, gaining traction especially in connection with initiatives such as HTR United, which promote shared frameworks for digital transcriptions that transcend the outdated disciplinary boundaries inherited from the 19th century. Platforms like *eScriptorium* further enhance this landscape by providing transcription environments where research collectives can collaborate and train custom handwriting text recognition (HTR) models. However, despite these advances, current standards remain under development and exhibit certain limitations. For example, TEI struggles to effectively manage authors' textual variants, while SegmentOnto does not handle well the complex entanglement of text regions within illustrations, tables, maps, or diagrams—all of which are critically important for scientific sources. Addressing these challenges is essential, but more crucially, the field lacks a user-friendly, accessible interface that would allow historians to experiment with these tools together and fully leverage their research potential. The creation of such an interface is a central aim of this project and justifies the need for a 18-month Research Engineer position in digital humanities.

The digital processing of texts, which enables the development of new methodologies combining both distant and close reading approaches, depends critically on the availability of robust OCR/HTR technologies for scientific sources. However, from historians' perspective, the OCR/HTR fields remain highly technical and deeply fragmented. While numerous models exist, they are often narrowly tailored to specific handwriting styles or, at best, to families of scripts, leaving historians grappling with significant technical barriers that are difficult to overcome. Addressing this challenge is a key objective of the project.

The project relies on a research team with a uniquely diverse portfolio of competencies. This collective is already operating informally through the collaboration of the ANR research project EIDA and VHS (run within the framework of SCAI at SU). It brings together historians (connected with the emerging IRHIST ASU's network) specializing in astronomy, mathematics, physics, music, natural history, and botany, focusing on sources in Chinese, Sanskrit, Persian, Arabic, Greek, Hebrew, French, English, Spanish and Latin, produced across Asia, Africa, and Europe from the 8th to the 19th centuries on both manuscript and printed corpora. Furthermore, thanks to a partnership with the Imagine team (LIGM, ENPC), the project benefits from exceptional expertise in applying computer vision methods to historical research.

In addition, team members from both the historical research group (Matthieu Husson, Alexandre Guilbaud, Stavros Lazaris, Eleonora Andriani) and the computer vision group (Matthieu Aubry, Raphaël Baena) have a proven track record in managing digital humanities initiatives and leading engineering teams within international research projects. This unique combination of skills positions the research team to pioneer novel computer vision and digital humanities approaches tailored for historians. Building on this analysis of the scientific landscape, the project will progress through the following steps.

In the domain of digital humanities, the recruited engineer, working in collaboration with the existing engineering teams of the EIDA and VHS projects, and in connection with SCAI, will first develop a data model designed to facilitate the processing and documentation of text regions in scientific sources. This model is expected to build upon existing standards such as SegmentOnto and will be aligned with the data models already in use within EIDA and VHS, which focus on scientific illustrations. Once this foundational task is completed the engineer will create a transcription interface. Initially, this tool will enable the team's historians to evaluate the performance of various HTR/OCR models. Over time, as the interface matures, it will propose automatic transcriptions and evolve into the project's central deliverable : a universal transcription application, freely available so that the community can use it on other corpora.

On the computer vision front, the team will design a 'meta' text recognition model that will adapt its behavior to the data, such as different alphabets or scripts. They will consider two complementary approaches. First, they will take stock of existing training datasets and models, and historians will annotate on our existing diverse set of documents which models perform well enough to be leveraged, and thus implicitly which data is relevant. This will enable to train a first network predicting for a given document which model should be used. Integrating such a model in our interface would significantly simplify the use of automatic text recognition models for large diverse datasets, such as ours. Second, we will develop a new approach, based on DTLR (Baena et al. NeurIPS 2024), which will jointly predict for each location in the text a character and a model to be used. Such a unified approach is more satisfactory from a theoretical perspective, and can be expected to improve performance, but it would also solve the practical challenge posed by texts with mixed fonts or languages, which is, for example, very common for early printed books.

The history team will play a pivotal role by contributing their expertise and access to primary sources, supporting both the computer vision and engineering teams. Additionally, they will conduct a methodological analysis of how the discursive and word-based elements of scientific sources were utilized by historical actors, offering critical insights into how these sources should be studied.

Transformative aspect

Highlight the originality, risk-taking, breakthrough and openness to society with regard to the research conducted in the field. (2 000 characters max, including spaces)

This research project aims to transform the field of the history of science by addressing significant methodological bottlenecks that hinder comprehensive engagement with historical scientific practices. Despite the discipline's expansion over recent decades, its reliance on outdated historiographical frameworks—rooted in 19th and early 20th-century practices—continues to perpetuate Eurocentric narratives and narrow disciplinary boundaries. The project proposes a novel approach that integrates recent advancements in computer vision and digital humanities to develop innovative digital tools. These tools will allow historians to analyze a wide range of manuscripts and printed sources across diverse chronological and geographical contexts, thereby literally producing a new methodological ground for the discipline. Central to this initiative is the exploration of the distinctive features of scientific texts, which often include non-discursive elements like illustrations, diagrams, and specialized terminologies all diversly connected by discursive word-based elements. The initiative aims to create a robust data model and a user-friendly automatic transcription interface that enhances historians' ability to leverage text recognition models. Concurrently, the project will advance the development of generalized computer vision techniques for text recognition, and a novel unified approach to joint model and character prediction. The project's interdisciplinary team, combining expertise from history and computer vision, is uniquely positioned to pioneer new methodologies. Crucially, the integration of humanities expertise from the outset—rather than as a retrospective addition—will ensure that the development of digital tools and methodologies is informed by historical context.

National and international positioning

This project advances the state of the art in digital humanities, history of science, and computer vision by addressing key challenges in the study of scientific texts. Recent developments, such as the Text Encoding Initiative (TEI) for digital publication, SegmentOnto for page segmentation, and eScriptorium for HTR, have improved textual data processing. Initiatives like HTR United have promoted shared frameworks for transcription, yet these tools face critical limitations with scientific sources. Texts in this domain often exhibit complex interactions between discursive elements (words and sentences) and non-discursive components (diagrams, maps, tables, etc.), which existing standards inadequately support. TEI struggles with authorial variants, while SegmentOnto fails to manage the integration of text regions with non-discursive features. Additionally, OCR/ HTR technologies are highly specialized, tailored to narrow handwriting styles or specific scripts, leaving historians with technical barriers to work across diverse textual traditions. The interdisciplinary and international team of the project is uniquely positioned to tackle these limitations. It includes historians of science from diverse disciplines—astronomy, mathematics, music, and botany—working with sources in Chinese, Sanskrit, Persian, Arabic, Greek, Hebrew, and Latin from the 8th to 19th centuries. Their expertise is complemented by computer vision researchers experienced in developing algorithms for historical research and leading digital humanities experts. This blend of skills will facilitate the development of tools that bridge current gaps in scholarship. The project will create a robust data model to document complex text regions of scientific sources, develop a transcription interface for evaluating existing HTR/OCR models, and analyze discursive elements to provide methodological insights, significantly advancing the digital humanities in the history of science.

Details of the financial request

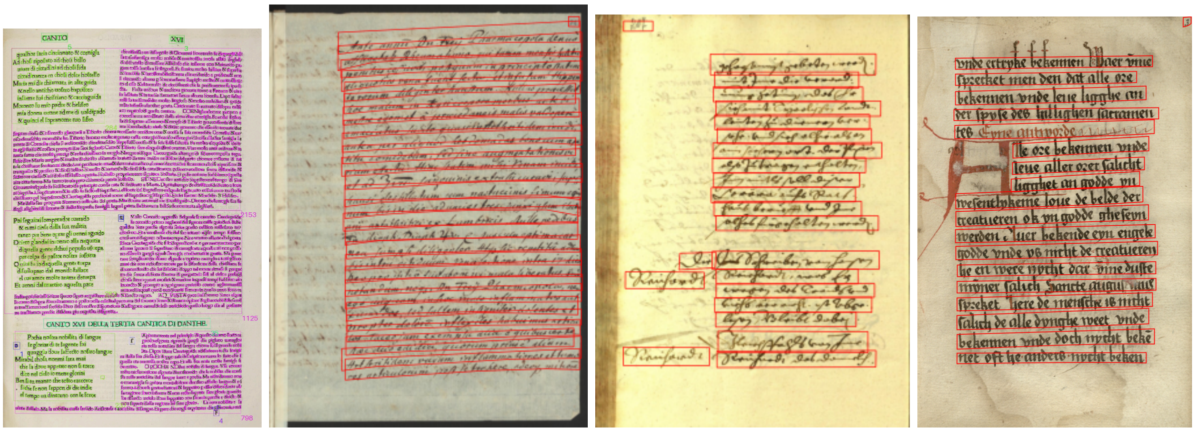
IR 18 month will work on the development of the digital platform and help manage the interaction between the computer vision and history part of the team 100 keuros

GPU will be necessary to run and train the HTR/OCR models 10 keuros

A third item in the budget is for small equipments and mission 5 keuros

Base paper: https://detection-based-text-line-recognition.github.io/

Proof of concept: We have an OCR model that works on complex pages and a line detection model that works in complex historical documents. Now the challenge is to have a transcription model that works on various



Text from Raphaël's CNRS proposal, do not re-use as is:
Current deep learning approaches for documents or epigraphs transcription are highly specialized: Optical Character Recognition (OCR) focuses on printed text, Handwritten Text Recognition (HTR) is restricted to handwritten documents, and epigraphic symbol recognition has been explored only through a few limited methods. Moreover, the previous approaches typically rely on domain-specific training, resulting in systems that are narrowly applicable to specific datasets with annotations. This results in poor generalization when applied to new scripts, languages, or symbol types.

Vision/HTR et Histoire

plusieurs models comment choisir le bon?
Manque DB, manque aussi interfaces pour que les historiens annotent.

Histoire
Outils pour charger un corpus et le transcrire automatiquement quelques soit la langue et son alphabet.
D'abord imprimé, puis manuscrits latins, grecs, arabes (?), Hebreu (?)
Moyen-Âge et période moderne

On dispose déjà d'outil de détections de lignes.
De modèles d'HTR

Enjeu : entrainer un outil à choisir le bon modèle pour transcrire efficacement → développement d'une plateforme d'annotation pour l'entrainement
Une équipe d'historiens pour annoter les transcriptions et entrainer l'outil

Livrable : une application (logiciel libre) utilisable par tous les médiévistes et modernistes de France et de Navarre :-)

Le développement de l'intelligence artificielle et des méthodes d'apprentissage ont très fortement et rapidement fait évolué, ces dernières années, les outils à la disposition des historiens pour travailler sur leurs corpus, qu'ils soient non seulement imprimés, mais également manuscrits. Des applications permettent aujourd'hui de réaliser des OCR ou des HRT de qualité de plus en plus hautes, les HTR reposant eux-mêmes sur des repérages des zones textuelles de plus en plus précis. Ces outils reposent sur de nombreux modèles qui ont pu être entraînés sur une très grande diversité de textes, relevant de différentes époques, cultures, dans différentes langues, sur différents supports : des applications telles que e-Scriptorium permettent à n'importe quel chercheur d'entraîner ces modèles sur le corpus de son choix et l'y appliquer pour obtenir une transcription.
Si l'on regarde l'ensemble des outils et techniques à notre disposition, il manque cependant encore un outil important, susceptible de modifier de façon radicale les conditions de travail et pratique de recherche des chercheurs en humanités travaillant sur les corpus manuscrits et imprimés : un outil de transcription universelle multi-lingue adapté aux corpus SHS, capable de mobiliser automatiquement l'ensemble des outils nécessaires, de l'identification des zones de texte à la sélection du modèle de transcription adapté à la langue ou aux langues du corpus pour en réaliser une transcription de la meilleure qualité possible. Un outil donc tout-en-un qui permettra aux nombreux chercheurs encore peu à l'aise avec les outils numériques, technologies de l'IA, avec l'entraînement de modèles, de disposer d'une application de transcription automatique en accès libre capable de s'adapter à la spécificité de chaque corpus. Un tel outil de transcription universelle des corpus et manuscrits anciens permettra à la fois de faciliter l'ensemble des travaux d'édition critique ou de data mining susceptible d'être réalisés sur ces textes, et il contribuera à réduire de façon significative le gap encore sensible existant entre les chercheurs en SHS et les outils issus des sciences des données.

benchmark existing HTR and OCR models using scientific sources, leveraging the historians' expertise to refine this process. Based on these benchmarks, they will train an algorithm capable of autonomously selecting the optimal transcription model for a specific library when confronted with new sources. Furthermore, as demonstrated in a published proof of concept, the computer vision group will develop an innovative, generalized approach to HTR and OCR rooted in character identification principles.