PRET19: Automatic Recognition and Indexing of Handwritten Loan Registers from 19th Century Parisian Universities

Léa Périssier, Viera Rebolledo-Dhuin, Marie-Thérèse Petiot, Yoann Schneider, Christopher Kermorvant

PRET19 : Automatic Recognition and Indexing of Handwritten Loan Registers from 19th Century Parisian Universities

Léa Périssier[1][0000−0001−5409−3052], Viera Rebolledo-Dhuin[2][0000−0002−4903−1372], Marie-Thérèse Petiot[1][0000−0003−2637−2139], Yoann Schneider[3], and Christopher Kermorvant[3][0000−0002−7508−4080]

[1] Bibliothèque interuniversitaire de la Sorbonne, Paris, France
[2] Centre de recherche en histoire européenne comparée, Université Paris-Est Créteil, France
[3] TEKLIA, Paris, France

Résumé The PRET19 project aims to carry out a comprehensive analysis of the nineteenth-century loan registers of several Parisian university libraries. These historical documents provide invaluable insights into the circulation of books and the intellectual engagement of the academic community during a transformative period. By reconstructing the relationships and trends between borrowers, the project offers a unique perspective on the intellectual landscape of Parisian universities. In this first phase, the registers come from three different university libraries and exhibit a great diversity in layout, handwriting and content, which poses significant challenges for data processing. To address these challenges, we developed a document processing workflow that effectively combines automatic handwritten text recognition (HTR) with manual processing and validation. In addition, we provide a detailed description of the database, which is designed to comprehensively model the information extracted from the registries, ensuring that the data is structured in a way that adequately responds to anticipated research requests. Once completed, the processed data will be from winter 2024 accessible *via* a dedicated website, providing a comprehensive digital resource for historical research.

Keywords: Librairies Loans Registers · Paris 19th Century · HTR · Deep Learning

1 Introduction

In 2020, to mark the 250th anniversary of the opening of the *Bibliothèque interuniversitaire de la Sorbonne* (BIS) to the public, a program was launched to digitize and study the archives and history of the library [6], which also forms part of the major thrust of the BIS's heritage policy, which is the history of the University of Paris [4]. It is within this framework that BIS is supporting

4. https://www.bis-sorbonne.fr/biu/spip.php?rubrique2

several projects, in particular ES LETTRES (2020-2023) [5] and PRET19(2022-2024) [6]. PRET19 (*Projet de Répertoire des Emprunteurs et Titres empruntés au XIX^e siècle à l'université*) focuses on the 19th century Parisian academic audience and its borrowing activity in three libraries of the Latin Quarter : the library of the Sorbonne , of the École normale supérieure and the Sainte-Geneviève Library. Digitization and scientific valorization of the loan registers aim at facilitating the exploration of those rather confidential manuscript sources, providing materials for research into the history of university libraries, their role in the circulation of scholarly books, and the borrowing practices of the academic audience, in the context of the transformation of academic work and disciplines during that time period. Its goal is to provide researchers with access to the digitized registers of the three libraries via a single site, making it possible to consult them and to search their contents by date, by borrower's name, and by various criteria resulting from the data enhancement regarding the borrowers. In a second phase, it may include the identification and enrichment of bibliographic data, allowing for a search by borrowed document. The project is designed to be open to other libraries sharing the same type of sources and audiences, for interoperability and partnerships.

The PRET19 project relies on a scientific committee of researchers and library curators from various disciplines in the social and human sciences : history, literature, sociology, history of mathematics, etc., whose many interests help to guide the work of creating the database.The projet s fully involved in the current renewal of studies of library loan registers at a time when Handwritten text recognition (HTR) offers offers new perspectives to researchers.

As Emmanuelle Chapron points out in an article reviewing the historiography on this subject up to 2021, loan registers have been the subject of several studies since the beginning of the twentieth century. Literary historians have used them to search for traces of "famous readers". The studies are then organised in "clusters" around a few libraries or a few readers, circulating from one library to another. From the 1980s onwards, as part of a social history of reading based on quantitative methods, library loan registers were used, initially in Germany, but soon in France, the United Kingdom and the United States, to investigate the penetration of books into the non-elite strata, if it is not to analyse the evolution of literary tastes, the progress of literacy or the penetration of Enlightenment ideas. A number of studies have also focused on communities of readers, in particular foreign readers and, less frequently, women readers. All these studies are based on two assumptions, which Emmanuelle Chapron invites us to go beyond : one equates borrowing books with reading them, and the other confuses the mass of borrowers with the public as a whole [9].

More recently, new approaches have emerged. Some of these are renewing research on "famous readers" in the sense of a history of intellectual work, shedding light on the genetics of works and the circulation of ideas from one writer to another. These include the work of Matthieu Béra and Nicolas Sembel, who

5. https://www.collexpersee.eu/projet/es-lettres/
6. https://www.collexpersee.eu/projet/pret19/

analyse the borrowings of Emile Durkheim and his nephew, Marcel Mauss, as well as that of Hippolyte Taine [22,21,1,2]. Emmanuelle Chapron, for her part, has devoted several publications to Bernard de Montfaucon and, in 2024, to Jean-François Séguier, in order to grasp the gesture of borrowing and its implications for the organization of intellectual work [8,10]. Other recent approaches shift their focus from the history of reading to the history of libraries. Bruno Blasselle and Ségolène Blettner are interested in the audiences and borrowers of the *Bibliothèque nationale de France* (BnF) [4,5], and since 1994 have been developing a research project on the "Anciens registres de prêt" of the BnF's Département des imprimés [7].

In other words, loan registers, already used by historians of books, libraries and literature at the beginning of the 20th century, are now a privileged object for the history of knowledge, taken in a broad sense. PRET19 has taken the gamble of using machine learning approaches to explore a body of work that exploded in the 19th century and we hope that this project will attract other libraries since the project aims at developing a process for all similar documents, with a view to creating data that can be consulted on a larger scale via an interoperable site.

The field of handwriting recognition has undergone significant advancements over the past two decades. The foundation for efficient handwriting recognition technology, particularly those based on statistical models, was established in the beginning of the 2000s [27]. However, the widespread adoption of this technology within the humanities has gained momentum more recently. This surge in adoption can be attributed primarily to advancements in deep learning technologies [13] and the emergence of accessible platforms that facilitate the training and application of HTR models, such as Transkribus [18], eScriptorium [20], and Arkindex [26].

In digital humanities projects, the goal often extends beyond simple transcription of handwritten documents. These projects typically aim to create databases of the information contained within these documents, such as registers [23] or tables [7]. Consequently, a mere transcription is insufficient, and a comprehensive workflow is required, which should include layout analysis, text recognition, information extraction and the crucial step of linking extracted information to existing indexes.

The objective of this paper is to explain the implementation of the procedure, from the digitization of the various sources and their automatic processing by HTR to the development of a single database facilitating research into borrowers and their loans within Parisian university libraries. This database could be extended to other similar projects.

Firstly, we give an overview of the sources, present the theoretical objectives of the project and describe the modelling of the data, with the aim of creating a single database that will be publicly accessible online. Secondly, we will describe the implementation of the complete workflow to automatically extract the information from the scanned documents to populate the database. This database will be built according to the FAIR principles.

7. http://http://comitehistoire.bnf.fr/recherches_en_cours

2 Loan registers for Parisian libraries in the 19th century : description of the corpus

The project is based on the corpus of three university libraries localised in the Latin Quarter : the *Bibliothèque interuniversitaire de la Sorbonne* (BIS), the *Bibliothèque de l'École normale supérieure de la rue d'Ulm* (ENS) and the *Bibliothèque Sainte-Geneviève* (BSG). Each of these institutions has its own implicit or explicit lending rules, as well as its own corpus, varying in quantity and quality, particularly in terms of writing

The series kept begin in 1811 for the Sorbonne and the École normale and from the year XI (1802-1803) for the BSG. The use of registers to manage external loans was abandoned in 1905 at the Sorbonne and in 1928 at the ENS. This explains the date range chosen for the project, namely 1811-1905.

At the Sorbonne, not only professors in the faculties and members of the major academic institutions (from Michelet to Bergson, not forgetting Augustin Thierry, Cousin, Ravaisson, Fustel de Coulanges, Duruy and Lavisse), but also lycée teachers and students at the École normale were able to borrow books. Loans were gradually opened up to other students from the 1870s onwards.

The corpus thus forms a set of approximately 39.000 views [8] of various contents and formats as detailed in Section 4.1.

From the point of view of the collections loaned, the disciplines concerned are numerous. They correspond to the thematic coverage of the various libraries at the time and evolved over the century. The humanities dominate the three partner libraries : ancient and modern literature and philology, history, geography, philosophy, moral and political sciences, anthropology, psychology and sociology. Theology, science and technology and law are also represented, particularly at the *Bibliothèque Sainte-Geneviève*, which preserves and develops encyclopaedic collections, and at the Sorbonne library, which serves three faculties : literature, science and theology. However, books on physical sciences and mathematics were also borrowed [9].

Focusing on the Parisian university public and its borrowing activities, the project aims to contribute to research into the history of university libraries and their role in the circulation of scholarly books, as well as the study of the scholarly public and its working practices. The identification of borrowers includes a strong prosopographical dimension, which is be able to draw on the many repositories and works already carried out in this area.

8. By "view", we mean a non-blank-page.

9. See the lecture on borrowing and borrowers of mathematics by N. Verber and V. Rebolledo-Dhuin at the SFHST congress, 19 April 2023 : https://hal.science/hal-04452449

TABLE 1 – Description of the loan register corpus in the three librairies of the Latin Quarter

Lib.	Nb of vol.	Types of Loan registers	Nb of views	1791-1800	1801-1810	1811-1820	1821-1830	1831-1840	1841-1850	1851-1860	1861-1870	1871-1880	1881-1890	1891-1900	1901-1910	1911-...
BIS	19	Registers of borrowers	6 540			1811									1904	
	12	Registers of authors	–				1824								1905	
	20	Daily registers	–						1845					1893		
	10	Others (ILL, statistics...)	–						1845							1945
ENS	50	Registers of borrowers	24 052			1812									1901	
	65	Registers of authors	–								1868					1928
BSG	6	Registers of borrowers	2 605	1800/1809/1812*								1872				
	1	Registers of authors	–					1838				1871				
	18	Daily registers	6 414						1846			1873			1904	

* the first Borrowers register of the BSG is widely incomplete, if it begins in 1800, it records loans only for these three years.
In grey, unprocessed digitized data, in color (in keeping with library logos), manually or automatically transcribed data.

Lib.	Registers of borrowers	Registers of authors	Daily registers
BIS	**Borrower(B)**** : Name, Adress, Occupation, Institution, Permission, Recommandation ; **Loan(L)** : dates (**loan and return –L&R**), Short title, **Registration nb (NUM)**, nb of vol., **person involved (PI)** ; Alpa. index of **B**)	**L** : Author (Alpha. order), Short title , **NUM**, nb of vol., **B**'s Name, dates (**L&R**)	**Date of L** (Chrono. order), **B**'s Name, Author, Short title, **NUM**, nb of vol., dates (**L&R**)
ENS	**B** : Name, status in the institution, **Signature (Sign.)** ; **L** : dates (**L&R**), Short title, **NUM**, **PI**	**L** : Author (Alphab. order) , Short title , **NUM**, nb of vol., **B**'s Name, dates (**L&R**)	
BSG	**B** : Name (Alphab. order), Adress, Occupation, Institution, Recommandation, **L** : Short title, **NUM**, dates (**L&R**), **PI**	**L** : Author (Alphab. order) , Short title, **NUM**, **B**'s Name, dates (**L&R**)	**Date of L** (Chrono. order), **B**'s name and **sign.**, Librarian's name or **sign.**, Author, Short title, **NUM**, nb of vol., **PI**, Date of return

**We bold abbreviated entries & underline the type of entry in each register

In addition to the list of works borrowed, which provides an insight into the history of the academic disciplines that were in the process of being established and institutionalised at the time, there are many other developments that could be envisaged : sketching out a topography of documentary practices in the Latin Quarter, through the movement of borrowers from one library to another ; exploiting borrowers' home addresses recorded in the registers, for example, by probing movements as the *cursus honorum* evolved. In addition to its documentary dimension, relating to the production, circulation and reception of works, the project is intended to be part of the history of science and the urban history of the capital.

In other words, the digitisation of loan registers and the accompanying tool are intended as a means to help reconstruct the intellectual dynamics, defined as encounters between individual trajectories and institutional constraints and opportunities, at the heart of various "ordres matériels du savoir" as understood by Christian Jacob, Françoise Waquet and Jean-François Bert and Jérôme Lamy [16,17,28,3]. Finally, in this same perspective of a material and social history of knowledge, these resources can provide elements for understanding the operation and development of this type of service (that of lending), between privilege and massification, the mechanisms put in place to manage the growing number of loans and control returns, and the mode of relationship that this service establishes between the library institution and the public authorised to borrow. The aim is to examine all aspects of these pragmatic writings of everyday life linking two parties, an institution and a borrower, and the mechanisms of trust that they presuppose and activate.

3 Building a Relational Database for Nineteenth-Century Library Loan Registers

3.1 Principles of the data modeling

The relational data modelling for the project was conducted prior to the automatic processing of the corpus. The primary objective was to identify the information contained within the loan registers in order to accurately represent it in a conceptual data model. This included a detailed exploration of the corpus from each library. Given the heterogeneity of the project's corpus, nearly all registers were examined in order to identify their unique features. In order to gain a deeper understanding of the context in which the lending practices occurred, the partner libraries were engaged to provide insights into the historical circumstances surrounding the era in question.

Three principal objectives guided this comprehensive examination of the sources. Firstly, the data model created needed to be precise enough to capture the full complexity of the content found in the registers, yet sufficiently general to be applicable across the three project libraries and potentially other libraries with similar collections.

Secondly, each piece of information within the database must be linked to the specific page of the register from which it was extracted. This linkage ensures that

visitors can verify each piece of information by accessing the historical source directly on the future website.

Finally, the model was designed not only to reflect the content of the registers, but also to allow for data enrichment. The work presented in this paper involved the identification of borrowers whose first names are rarely recorded in the registers. In addition, the historical geolocation of the addresses was determined by querying the API of the SoDUCo Historical Geocoder [12]. Future work will focus on semi-automatic identification of borrowed titles, which are often only briefly described in the registers.

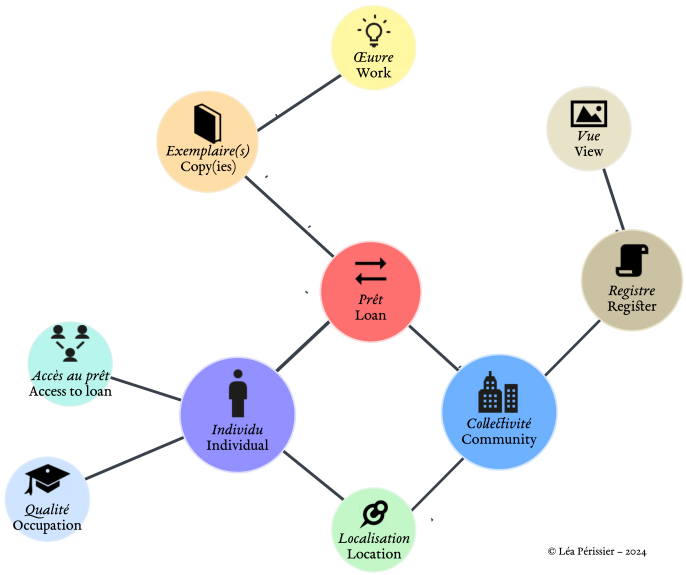


FIGURE 1 – Synthetic data model used in the PRET19 project

3.2 Description of the data model

The complete data model is shown in Figure 6.1. The model comprises ten entities, categorised into two distinct types. Entities in pink represent information sourced from original documents, accompanied by attributes intended for enrichment. Entities in green exclusively contain information drawn from these sources. In the model, attributes dedicated to enrichment are located below the dashed lines. The data model can be divided into three subsets. The first subset comprises the Register (*Registre*) and View (*Vue*) entities, relating to the project's source documents. The Register entity represents a loan register whose details are stored in the database. The View entity, positioned centrally in the model, corresponds to a scanned page of a register. All other entities in the model

are linked to the View, facilitating access to the source, as previously mentioned. A second subset is formed by the entities Person (*Personne*), Collectivity (*Collectivité*), Localization (*Localisation*), Quality (*Qualité*), and Loan Access (*Accès au prêt*). This part of the model is concerned with information pertaining to the actors mentioned in the registers. Any individual mentioned, regardless of their role, is represented by the Person entity. The Collectivity entity refers to a legal entity, either public or private. The Quality entity pertains to any information about a person's role as it appears in the source, including title, profession, and status. Any address present in the sources is represented by the Localization entity. For instance, attributes for this entity dedicated to enrichment include geographic coordinates, the date used to query the historical API of SoDUCo, and the restored city (when it is not mentioned in the sources). Finally, the Loan Access entity corresponds to a particular mention of loan access for a borrower, which could include a letter of recommendation, certificates, etc.

The last subset consists of the entities Loan (*Prêt*), Copies (*Exemplaire(s)*), and Work (*Œuvre*). It encompasses all information regarding the transaction between a borrower and a borrowed title. The Copies entity represents the physical items borrowed by an individual. This entity is inspired by the *Item* entity from the Functional Requirements for Bibliographic Records [15]. Developed by an expert group from International Federation of Library Associations and Institutions [10], this model provides a conceptual modelling of information contained in bibliographic records. A copy is identified by its shelf-mark, which may refer to multiple volumes.

The Loan entity relates to transaction details between a borrower and one or more copies, corresponding to a line in a register : loan date, return date, special circumstances, etc. An example of adapting the model to meet the specificities of the three libraries is demonstrated through the Involve (*Impliquer*) association between the Loan and Person entities. This association allows adding a person or entity mentioned alongside a loan, whether their role in the transaction is defined or not. Depending on the library, this relationship can cover various realities. For ENS, it typically involves a third party for whom a borrower makes a loan. At BSG, it may involve persons for whom the librarian records loans under their own name, but it is challenging to interpret the data at this stage of work.

Finally, the Work entity refers to the abstract creation to which the Copies entity is connected. It is a conceptual entity because the borrowed titles are usually described only by their title and author, without further details about the publication (publisher, date, etc.).

The data model for the project was implemented using the Heurist framework [11], an open-source relational database management system particularly well-suited to projects in the humanities and social sciences. We utilize an instance of Heurist hosted by IR* Huma-Num, which manages the installation and ongoing maintenance of the software, thereby ensuring the long-term preservation and stability of the data. Heurist is compatible with the International

10. https://www.ifla.org/
11. https://heuristnetwork.org/.

Image Interoperability Framework (IIIF), which the project's libraries utilise to access images. This compatibility is of the utmost importance for the seamless integration of image viewing within the database, facilitated by the customisable web interface that Heurist provides. This configuration enables direct and user-friendly access to the digital resources managed within the system.

4 Semi-automatic indexing of handwritten registers

In the field of historical document processing, the choice between manual and automated methods is crucial and is primarily determined by the complexity and volume of the documents involved. Documents with unique or intricate features often require manual processing to ensure the accuracy and fidelity of the extracted information. Automated processing, on the other hand, is preferred for larger collections where the uniformity and predictability of document types allows for efficient processing at scale.

The decision between these processing methods depends on the trade-offs between accuracy, cost and scalability. Projects often use a strategic mix of both, automating bulk processing and manually processing documents identified as outliers or of particular importance to ensure accuracy where required. This balanced approach optimises both the use of resources and the quality of the output.

4.1 Analysis of the corpus complexity

The corpus, summarized in Table 4, is more complex to process than it might seem, given its mass and diversity. The difficulty lies in the diversity of layout, handwritings and the erasures that can be added to indicate the return of a loan, as shown on Figure 2.

On the basis of this detailed analysis of the structural characteristics and the degree of uniformity in the presentation of the content of each register, we have decided on the strategy for managing their processing according to the following rules :

For registers with a homogeneous layout and standardised information, in particular those without excessive erasures and with a significant number of pages, we use an automated processing approach. Specifically, a dedicated model is trained for each of the main types of indexes identified within this category. These models are designed to handle the bulk of the processing workload, with an emphasis on efficiency and consistency in data extraction. To ensure accuracy, results that the model predicts with low confidence are manually validated. This dual approach exploits the speed and scalability of automated processing, while maintaining the quality assurance provided by manual review.

TABLE 4 – Content and processing of borrower registers for the 3 partner libraries (BIS, ENS, BSG)

	Dates	**Nb of vol.**	**Nb of views**	**Layout complexity**	**Writing & Information Complexity**	**Processing**
BIS	1812-1850	10	1.907	Non-standardized, which can vary within the same register	Presence of scribbles in the oldest registers + Few details (many blank pages)	Manual
	1850-1904	8	4.633	Standardized, each page contains 6 columns	Standardized information on the borrower (occupation or title, address) and the dates of borrowing and returning, and remarks + significant number of pages	Automatic
ENS	1813-1849	3	843	Non-standardized, unique to each registry, with one of them having two layouts	Low quantity (many blank pages) + heterogeneous layouts	Manual
	1863-1900	45	23.179	Rather standardized layout. A page generally appears as follows : the borrower's name at the top of the page followed by his loans. Predominantly one borrower per page.	Large number of pages (+ automatic exclusion of the many blank pages of these registers : blotting papers)	Automatic
BSG	1800-1872	6	2.605	Random, unique to each registry	Low amount of information (many blank pages) + significant scribbles at the beginning of the century that hinder readability	Manual
	1873-1904	13	6.414	Standardized content. Each register contains two types of pages : "Exit" pages containing information on the borrowing of a title and "Return" pages containing information on the restitution of a title	Standardized + Large number of pages	Automatic
Total		**82**	**39.611**			

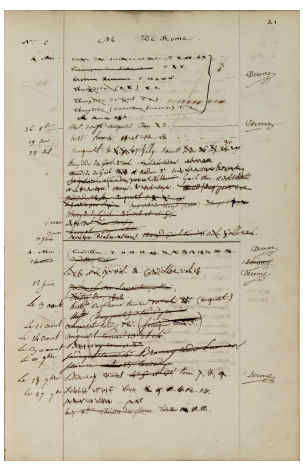
Conversely, indexes characterised by heterogeneous layouts, non-standardised information presentation, frequent erasures and a relatively small number of pages are processed entirely manually. This decision stems from the recognition that automated systems may struggle with the irregularities and unique characteristics of these documents. Manual processing, although more time-consuming, allows for a nuanced understanding and personalised treatment of each page, which is critical to maintaining the integrity and accuracy of the data extracted from these complex sources.

This individualised approach ensures that each register is processed in a manner best suited to its specific characteristics, maximising both efficiency and data fidelity across the corpus.

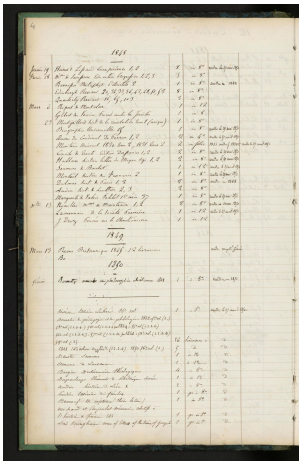
4.2 Description of the workflow

Figure 3 illustrates the workflow adopted by our project for extracting and processing borrower information from the historical loan registers of three libraries. The first stage of the pipeline involves a classifier, a component that detects the presence of relevant information on each page. Its primary function is to filter out pages without borrower data (cover pages, blank pages, index pages, etc) thereby simplifying the processing of large collections by focusing on content-rich pages and avoiding hallucination of the recognition model. Pages confirmed to contain borrower information are passed to the segmenter. The goal here is to isolate the specific area of the document that contains borrower information. This segmentation facilitates targeted analysis by grouping together all information relating to a specific individual, thereby increasing the efficiency of subsequent recognition. The workflow then moves to the Recognizer. This module uses a text recognition model that is capable of extracting key-value pairs from handwritten text. The recogniser is designed to recognise and type these handwritten entries, converting them into structured data that can be further processed. The final phase is indexing, where the extracted information is matched against existing referential databases. This step is critical in integrating the processed data into searchable and analysable datasets, making it accessible for historical research and analysis.

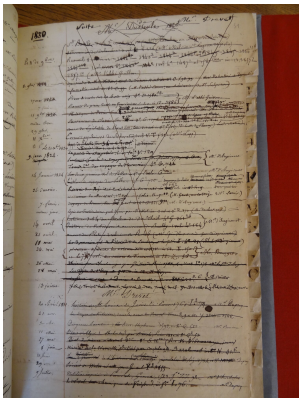
To implement and manage this workflow, the project uses two main tools. For managing and processing the documents, we relied on Arkindex, an open source document processing platform [26] that supports the processing of large document collections through complex machine learning and deep learning pipelines. Arkindex natively supports the International Image Interoperability Framework (IIIF) and connects directly to the library's to the library's IIIF API endpoints. All algorithms are built into Arkindex, simplifying data handling and processing. For the generation of ground truth data and for validating low confidence predictions, we used the open-source annotation tool Callico [19]. The following paragraphs provide a technical description of each component of the processing workflow.



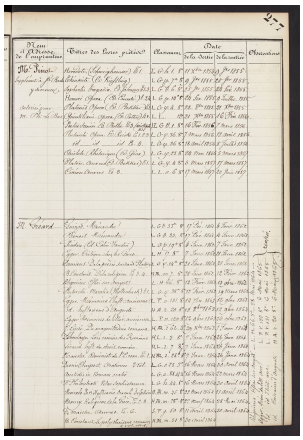
(a) BIS MS 409



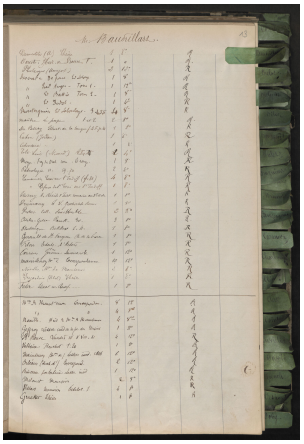
(b) ENS AB/1



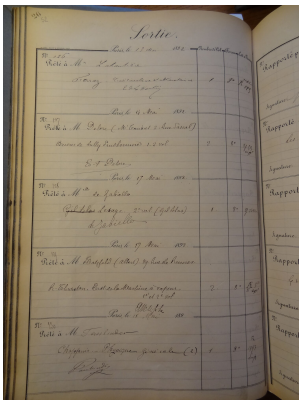
(c) BSG 2006/01/247



(d) BIS AAM 35



(e) ENS AB/28



(f) BSG 2006-01/220

FIGURE 2 – This figure shows a sample page from different registers of the BIS, ENS and BSG libraries to illustrate the variability in layout and information complexity. The first row shows volumes that are either too complex or too few to be processed automatically ; these were processed manually using the Callico platform interface. The second row shows regular volumes, available in large quantities, which were processed using dedicated automated models and then manually validated to ensure accuracy.

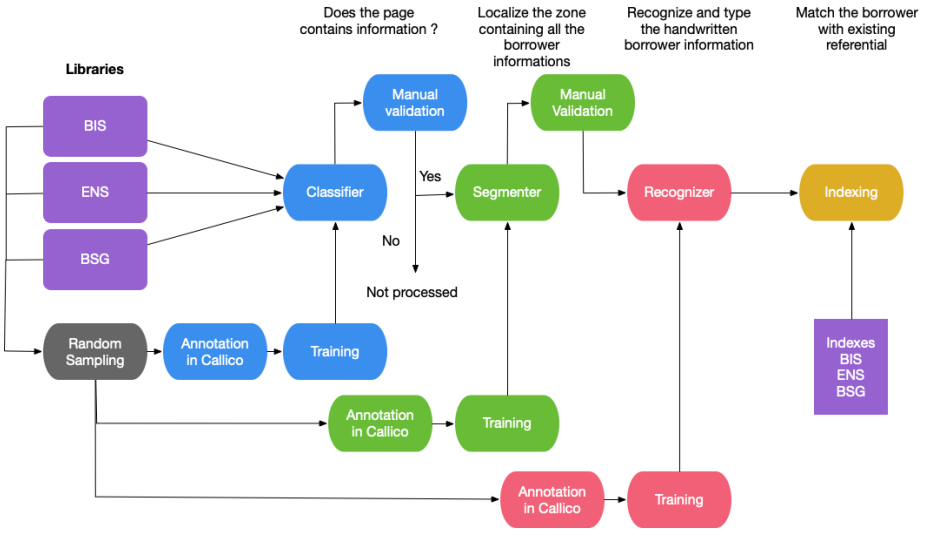


FIGURE 3 – Semi-automated document processing workflow. Pages are classified for relevant borrower information, then segmented and recognized. Recognized information are matched against reference indexes. The workflow includes a feedback loop using manual corrections to enhance recognizer accuracy.

Classifier : For the page classification, we trained the classifier using the YOLOv8 library [12] on a set of 300 images randomly selected from all the images in the three libraries' corpus. YOLOv8 was chosen because it is open-source, and it typically achieves very good performance on document classification tasks using only the visual aspect of the document (the content is not used) with a reduced set of annotated examples.

The sample images were manually annotated using the classification mode in Callico, then resized to 224x224 pixels and segmented into training (80%), validation (10%), and test (10%) sets. The model was fine-tuned from the YOLOv8x-cls model, with a batch size of 8 on two RTX 3090 GPUs. The classifier achieved perfect precision and recall rates of 1 on the test set. To ensure the utmost reliability, manual validation was undertaken for the full corpus of 21,000 pages. This process was completed by a single annotator in a span of 7 hours, using the batch validation view in Arkindex.

Segmenter : The role of the segmenter in our pipeline is to accurately locate zones of borrower information within the digitised pages. We used the YOLOv8 segmentation model for this task. A first sample of 300 pages was randomly selected from all the pages containing borrower information, as they were all manually validated during the classification step. This sample was annotated

12. https://docs.ultralytics.com/tasks/

using Callico's segmentation mode and divided into 80% for training, 10% for validation and 10% for testing. A YOLOv8 model was fine-tuned on this sample, starting from the model pre-trained on the COCO dataset provided by the library, in detection mode for 600 epochs using early stopping.

On initial evaluation, the model demonstrated performance on the test set with a precision (BoxP) of 0.954, a recall (BoxR) of 0.978, and a mean average precision at a 50% intersection over the union threshold (Box mAP50) of 0.978.

Despite these high metrics, a detailed error analysis was performed, which identified 103 pages where the model's performance was suboptimal. These pages were specifically annotated and added to the training set to improve the model's ability to handle complex cases. After retraining, the new model showed a Box Precision (BoxP) of 0.818, Box Recall (BoxR) of 0.544, and a Mean Average Precision at 50% IoU (Box mAP50) of 0.629 on these challenging images. In contrast, the model trained on pages from the BSG library, using a smaller set of 100 images, achieved near perfect metrics with a BoxP of 0.999, a BoxR of 1 and a Box mAP50 of 0.995.

Handwritten information extraction : A Document Attention Network (DAN) [11,25] was used to extract handwritten borrower information from historical documents. This model integrates Handwritten Text Recognition (HTR) with Named Entity Recognition (NER) to simultaneously identify and classify textual content, focusing on relevant borrower details while excluding irrelevant data such as book titles. The model was pre-trained on the French RIMES database[14] and further fine-tuned on 583 "emprunteur" zones extracted from our corpus, divided into 482 training, 55 validation and 46 test zones. The sample was annotated using the entity form mode in Callico, which allows all the information related to a borrower zone to be entered using a simple interface that presents the image and a form. The form contained 79 fields grouped into 7 sections corresponding to the information required to complete the data model. Not all information was available for each borrower, and in many cases the fields were left blank in the annotation interface.

Training was performed on an NVIDIA A100 GPU with 80 GB, running for 880 epochs with early stopping, using a batch size of 2. Images of the borrower zone were resized to a maximum of 1500 pixels wide and 3000 pixels high. The model's results for the BIS/ENS dataset were relatively modest, with Precision, Recall, and F1 scores of 0.60, 0.50, and 0.54, respectively. For the BSG volumes, the model showed improved performance, with Precision at 0.71, Recall at 0.77, and an F1 score of 0.74. Despite these moderate overall results, the model was particularly effective in extracting borrower names, achieving an F1 score of 0.75 for BIS/ENS and 0.76 for BSG. This higher accuracy in name extraction is crucial for our project, as it sufficiently enables borrower identification within a referential database by integrating multiple pieces of information. The effectiveness of this approach in accurately identifying borrowers will be discussed in the next section on borrower identification.

In order to facilitate the validation of predictions, the confidence scores were calibrated using the technique of temperature scaling [24]. This adjustment renders the score more useful for the purpose of filtering out the most uncertain predictions, which are then subjected to manual validation in the Callico data validation interface.

Matching : The workflow incorporates four distinct referential indexes for matching recognised borrower information :

1. Multi-Library Index : A manually compiled collection from various historical sources, including EPHE's prosopographic dictionary, LARHRA's Faculty Professors (1808-1880) and Secondary Education Aggregates (1809-1960) databases, and the BIS-led 'ès lettres' project database.
2. BSG Library Index : A working file created by BSG staff, containing a non-exhaustive list of borrowers and staff from partially processed registers, reflecting the ongoing nature of the compilation.
3. ENS Library Index : Manually compiled from lists of borrowers' names on Calames, with contributions from ENS library staff, covering 45 of the 48 available registers.
4. BIS Library Index : An index of BIS borrowers created from index pages found in eight registers from 1850 to 1904. These pages were processed using a generic PyLaia HTR model[24], with predictions validated by BIS experts for accuracy.

Each referential index contains the following information on the borrowers : name, first name, and for most associated with ENS, the date of birth, date of death, and registry number.

The referential indexes are ingested into an Elasticsearch [13] search engine to facilitate robust and flexible searching. This method does not require the training of a machine learning model, but rather the definition of the rules used to define a match during the search. The search algorithm operates on all borrower zones where a name has been detected by the model. If a first name is also detected, it is included in the search parameters. The following search strategy and criteria are employed : fuzzy search is implemented for the borrower's last name to account for minor discrepancies ; exact search is applied to the register number for ENS registers, except in certain specified cases ; if the first name was recognized, the first letter of the first name must match exactly and the remaining characters, if more than one letter is predicted, are subject to a fuzzy search. Then the search is prioritised as follows : first search within the reference of the borrower's home library (with register restrictions for ENS), then search within the references without registry constraints for ENS (except for exceptions) and finally search across the references of other libraries.

The quantitative results for matching predicted borrowers to their referential databases across different libraries demonstrate high effectiveness in the

13. https://www.elastic.co/elasticsearch

automated matching process. In the ENS registers, 2,175 out of 2,661 predicted borrowers were matched, achieving an 81.7% success rate. For the BIS library, the algorithm demonstrated a slightly higher effectiveness, matching 8,379 out of 9,791 borrowers, resulting in an 85.6% match rate. Similarly, in the BSG registers, we identified 10,389 out of 12,691 predicted borrowers, which corresponds to an 81.9% match rate.

It is important to note that a match does not necessarily mean that the person has been correctly identified. The person's information may contain errors due to automatic transcription, and the rules may be too loose, allowing false matches. Systematic manual validation would be required to accurately assess the accuracy of the matches.

5 Conclusion and future work

The PRET19 project initiated an ambitious endeavour to analyse loan registers from several Parisian university libraries dating back to the 19th century. With over 25 members, the scientific committee brings a broad array of scientific perspectives and research objectives, significantly enriching the scope of the project. This diversity presents both an opportunity and a challenge, necessitating a comprehensive and precise modelling of data to adequately cater to all anticipated research inquiries.

In addition, the diversity of registers originating from different institutions and time periods presents a complex array of formats and content. To manage this effectively, we have developed a robust semi-automated processing workflow that combines deep learning with manual intervention. This approach ensures efficient processing and accuracy, using expert knowledge to capture nuanced and contextual information.

Finally, the public-facing website of the PRET19 project – avaible from winter 2024 [14] – will act as a comprehensive digital portal, allowing users to interactively explore records of different entities such as persons, collectives, transactions and sources using faceted search capabilities. Users will be able to view detailed data alongside original sources, comparing automatic work with manual transcription, or to supplement the database directly with manual entries – via validation bodies, controlled by which enable collaborative data entry ?. In addition, the site will include extensive documentation on 19th century lending practices, including regulations, statistics and historical context, enriching our understanding of library operations and interactions during this period.

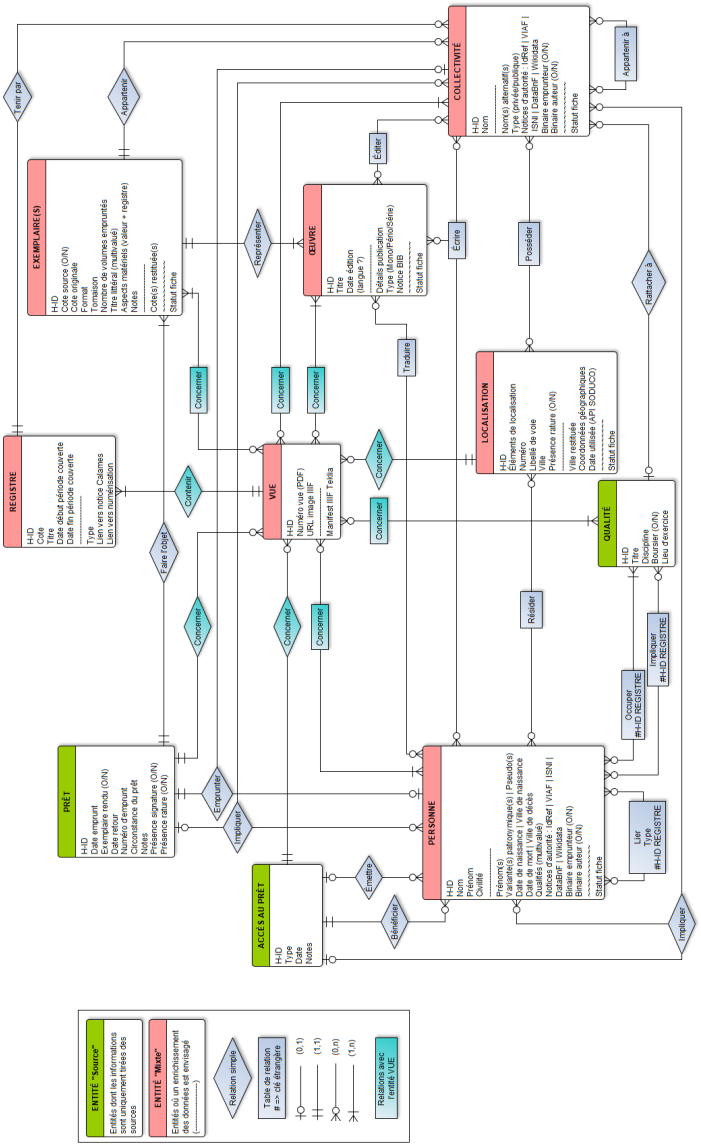
Acknowledgments : We thank the GIS CollEx-Persée for their financial support. The GIS CollEx-Persée, under the auspices of the French Ministry of Higher Education, Research, and Innovation (MESRI), has been instrumental in enhancing the accessibility of heritage documents for scientific research through a national network of library cooperation established in 2017.

14. At the following adress :https://pret19.bis-sorbonne.fr/, see too 6.2

6 Appendices

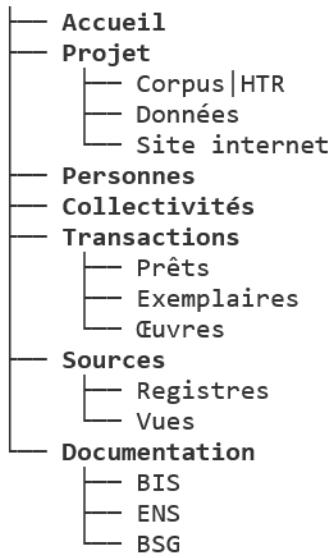
6.1 Data model detailled

Presentation of the data model used in the project, detailing the relational structure among various entities associated with historical library loan registers. Key entities such as *Prêt* (Loan), *Registre* (Register), *Vue* (View), *Exemplaire(s)* (Copy(ies)), *Œuvre* (Work), *Personne* (Individual), *Collectivité* (Community), *Localisation* (Location), and *Qualité* (Profession) are interconnected to represent both source-derived information and enriched metadata.



6.2 Structure of the website

The website (currently under construction, in private mode, end due to go public in winter 2024), includes static pages (documentation) and database query pages.



The first two tabs (*Accueil*, *Projet*) present information on the project, the corpus concerned and the context in which the data was created. The 3rd, 4th, 5th an 6th (*Personnes*, *Collectivités*, *Transactions*, *Sources*) are used to query the database. The "visitor" can search *via* facets and export the results in CSV or JSON formats, for example. The website's ergonomics – wich owes much to the *Scripta Manent* project [15], also produced under Heurist – feature pop-ups that allow you to navigate from a search in one or more related entities, without losing legibility. So, in the PRET19 project, pop-ups are used to display the sources (original register pages) in Mirador, along with the records in the database. Finally, the last tab (*Documentation*) - a static page - provides documentation on lending in each of the libraries throughout the 19th century : regulations, background information and statistics on the type of borrower according to origin (Faculty of Science or Literature), status (professor or student), etc.

15. https://heurist.huma-num.fr/ScriptaManent/web/13822/36204

Références

1. Béra, M. : Les emprunts de Durkheim dans les bibliothèques de l'École normale supérieure et de la Sorbonne, 1902–1917. Durkheimian Studies **22**(1), 3–46 (2016). https://doi.org/10.3167/ds.2016.220101, https://www.berghahnjournals.com/view/journals/durkheimian-studies/22/1/ds220101.xml
2. Béra, M. : Taine indiscipliné, mais discipliné. Les voies de l'assignation disciplinaire d'un auteur dans les bibliothèques. Les Études Sociales **174**(2), 115–150 (2021). https://doi.org/10.3917/etsoc.174.0115, https://www.cairn.info/revue-les-etudes-sociales-2021-2-page-115.htm
3. Bert, J.F., Lamy, J. : Voir les savoirs. Lieux, objets et gestes de la science. Anamosa, Paris (2021), https://journals.openedition.org/lectures/50810
4. Blasselle, B. : Les lecteurs de la Bibliothèque nationale au XIXe siècle. L'apport des registres de prêt. Les Études Sociales **166**(2), 69–88 (2017). https://doi.org/10.3917/etsoc.166.0069, https://www.cairn.info/revue-les-etudes-sociales-2017-2-page-69.htm
5. Blasselle, B., Blettner, S. : Lecteurs et emprunteurs à la Bibliothèque royale sous la monarchie de Juillet. Romantisme **177**(3), 8–19 (2017), https://www.cairn.info/revue-romantisme-2017-3-page-8.htm
6. Bobis, L., Noguès, B. (eds.) : La bibliothèque de la Sorbonne. 250 ans d'histoire au cœur de l'université. No. 87 in Histoire de la France aux XIX^{e} et XX^{e} siècles, Éditions de la Sorbonne, Paris (2021)
7. Boillet, M., Tarride, S., Schneider, Y., Abadie, B., Kesztenbaum, L., Kermorvant, C. : The socface project : Large-scale collection, processing, and analysis of a century of french censuses. In : Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) (2024)
8. Chapron, E. : Les bibliothèques de Bernard de Montfaucon. In : Krings, V., Jestaz, J. (eds.) L'antiquité expliquée et représentée en figures "de Bernard de Montfaucon. Histoire d'un livre. No. 19 in Scripta receptoria, Ausonius éditions, Bordeaux Pessac (2021)
9. Chapron, E. : Les registres de prêt des bibliothèques : De l'histoire de la lecture à l'histoire des bibliothèques. Francia **48**, 123–144 (2021). https://doi.org/10.11588/fr.2021.1.93925, https://journals.ub.uni-heidelberg.de/index.php/fr/article/view/93925
10. Chapron, E. : La vie dans les papiers. Jean-François Séguier (1703–1784). No. 3 in Heuristiques, Schwabe (2024)
11. Coquenet, D., Chatelain, C., Paquet, T. : DAN : a segmentation-free document attention network for handwritten document recognition. In : IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1–17. Institute of Electrical and Electronics Engineers (IEEE) (Jan 2023). https://doi.org/10.1109/tpami.2023.3235826
12. Cura, R., Dumenieu, B., Abadie, N., Costes, B., Perret, J., Gribaudi, M. : Historical collaborative geocoding. ISPRS International Journal of Geo-Information **7**(7), 262 (2018). https://doi.org/10.3390/ijgi7070262, https://doi.org/10.3390/ijgi7070262
13. Graves, A., Schmidhuber, J. : Offline handwriting recognition with multidimensional recurrent neural networks. In : Neural Information Processing Systems (2008), https://api.semanticscholar.org/CorpusID:639755

14. Grosicki, E., Carré, M., Geoffrois, E., Augustin, E., Preteux, F., Messina, R. : Rimes, complete (Mar 2024). https://doi.org/10.5281/zenodo.10812725, https://doi.org/10.5281/zenodo.10812725
15. International Federation of Library Associations and Institutions : Functional requirements for bibliographic records. Final report, IFLA Study Group on the Functional Requirements for Bibliographic Records (1997), approved by the Standing Committee of the IFLA Section on Cataloguing. Amended and corrected through February 2009
16. Jacob, C. (ed.) : Lieux de savoir. Tome 1 : Espaces et communautés. Albin Michel, Paris (2007)
17. Jacob, C. (ed.) : Lieux de savoir. Tome 2 : Les mains de l'intellect. Albin Michel, Paris (2010)
18. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G. : Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In : 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 04, pp. 19–24 (2017)
19. Kermorvant, C., Bardou, E., Blanco, M., Abadie, B. : Callico : a versatile open-source document image annotation platform. In : Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) (2024)
20. Kiessling, B., Tissot, R., Stokes, P., Stökl Ben Ezra, D. : eScriptorium : An open source platform for historical document analysis. In : 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19 (2019)
21. Sembel, N. : Les emprunts de Mauss à la bibliothèque universitaire de Bordeaux : la genèse d'une « imagination sociologique ». Durkheimian Studies **21**(1), 3–60 (2015). https://doi.org/10.3167/ds.2015.210101, http://berghahnjournals.com/view/journals/durkheimian-studies/21/1/ds210101.xml
22. Sembel, N., Béra, M. : Emprunts de Durkheim à la bibliothèque universitaire de Bordeaux : 1889-1902. Durkheimian Studies **19**(1), 49–71 (2013). https://doi.org/10.3167/ds.2013.190103, http://berghahnjournals.com/view/journals/durkheimian-studies/19/1/ds190103.xml
23. Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., Vézina, H., Kermorvant, C. : Large-scale genealogical information extraction from handwritten Quebec parish records. International Journal on Document Analysis and Recognition (IJDAR) **26**(3), 255–272 (Sep 2023). https://doi.org/10.1007/s10032-023-00427-w, https://doi.org/10.1007/s10032-023-00427-w
24. Tarride, S., Schneider, Y., Generali-Lince, M., Boillet, M., Abadie, B., Kermorvant, C. : Improving automatic text recognition with language models in the pylaia open-source library. In : Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) (2024)
25. Tarride, S., Boillet, M., Kermorvant, C. : Key-Value Information Extraction from Full Handwritten Pages. In : Document Analysis and Recognition - ICDAR 2023. pp. 185–204. Springer Nature Switzerland (Aug 2023). https://doi.org/10.1007/978-3-031-41679-8_11
26. TEKLIA : Arkindex : A document processing platform. https://doc.arkindex.org/ (2024), accessed : 2024-05-07

27. Vinciarelli, A., Bengio, S., Bunke, H. : Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 709–720 (jun 2004)
28. Waquet, F. : L'ordre matériel du savoir. Comment les savants travaillent (XVI^e^-XXI^e^ siècles). CNRS Éditions, Paris (2015), https://www.cnrseditions.fr/catalogue/histoire/lordre-materiel-du-savoir/