Knowledge Database for Historical Demography

Alain Bouju[1,*], Mickaël Coustaty[1], Thomas Roy[1], Isabelle Seguy[2], Arnaud Bringé[2], Pascal Chareille[3], Doriane Hare[4] and Florian Langele[4]

[1]*L3i, Avenue Michel Crépeau, La Rochelle, France*
[2]*INED, 9 cours des Humanités, Aubervilliers, France*
[3]*CeTHiS, 3 rue des Tanneurs, Tours, France*
[4]*CRM, 1, rue Victor Cousin, Paris, France*

Abstract

This paper introduces the architectural and semantic foundations of the DAI-CRéTDHI project, a national digital humanities initiative designed to extract, model, and interconnect multi-source French historical demographic data from the 16th century onward. Converting handwritten parish and civil registers into structured knowledge graphs requires overcoming significant material degradation, layout variability, complex administrative mutations, and massive data redundancy. Our pipeline addresses these challenges through a workflow. First, an automated document processing framework deploys a custom, YOLO-derived object detection model for record segmentation and leverages Teklia's reference Handwritten Text Recognition (HTR) engines for text extraction. Second, we present a streamlined ontology extending OWL-Time and GeoSPARQL that acts as a strategic conceptual pivot, connecting the institutional framework of the ANR Obardi project with the continuous geographic map of the ANR Communes initiative.

A core contribution of our approach is the systematic modeling of data provenance, which decouples transient historical assertions (PersonEvent) from global persistent identities. By populating our knowledge base with massive, crowdsourced datasets provided by Geneanet, we highlight and address a critical historiographical hurdle: the coexistence of multiple entries of the same person across individual life paths, and duplicate records generated by collaborative indexing overlaps and native archival redundancies (such as the dual preservation of local communal and court *greffe* registers, paired with decade-long summaries or *Tables Décennales*). To resolve these co-referent nodes without introducing graph-reasoning bottlenecks, we integrate Splink, an external probabilistic record linkage framework based on the Fellegi-Sunter methodology. Finally, we detail an ongoing benchmark framework—evaluating Apache Jena/Fuseki, Ontotext GraphDB, and a spatially enhanced version of Apache Marmotta over a PostgreSQL/PostGIS backend. This ensures scalable geo-historical querying over millions of triples.

Keywords

Parish register, Semantic Web for historical demography, Handwritten Text Recognition

1. Introduction

This paper introduces DAI-CRETDHI[1], a national project funded by the French National Research Agency (ANR). The project's primary objective is to manage and model comprehensive data on the French population at both individual and societal levels. Historical demography, which investigates the demographic evolution of societies, is a well-established field where researchers have traditionally relied on the manual analysis of registers and archives. However, this manual workflow is deeply labor-intensive and inherently constrained by the volume of documents that can be processed. With the advent of large-scale digitization, innovative infrastructure projects such as the Balsac database[2] have

SWODCH'26: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Workshop days of ISWC 2026, Bari, Italy

*Corresponding author.
alain.bouju@univ-lr.fr (A. Bouju); mickael.coustaty@univ-lr.fr (M. Coustaty); thomas.roy@univ-lr.fr (T. Roy); seguy@ined.fr (I. Seguy); bringe@ined.fr (A. Bringé); pascal.chareille@univ-tours.fr (P. Chareille); doriane.hare@cnrs.fr (D. Hare); florian.langele@sorbonne-universite.fr (F. Langele)
0000-0002-1599-9326 (A. Bouju); 0000-0002-0123-439X (M. Coustaty); 0009-0001-5877-2914 (T. Roy); 0000-0002-6833-570X (I. Seguy); 0009-0009-2680-1512 (A. Bringé); 0000-0002-9553-2442 (P. Chareille); 0009-0005-1483-7738 (D. Hare); 0009-0005-0376-4086 (F. Langele)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://dai-cretdhi.univ-lr.fr/
[2]https://balsac.uqac.ca/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4320/paper3.pdf

emerged to store and analyze extensive population datasets. Focused on Quebec's population from the 17th century to the present, Balsac has received continuous support since the 1980s, establishing itself as a major international reference. Similar demographic infrastructure models have been successfully deployed across Europe, including the Historical Sample of the Netherlands (IISH), the Swedish Historical Population Statistics (Umeå University), the Antwerp-Cor-Database, and large-scale population reconstruction initiatives such as Link-Lives. While these initiatives have introduced major innovations in automated text recognition and data storage optimization, significant semantic challenges remain. In particular, the automated recognition, disambiguation, and entity-linking of named entities within heterogeneous heritage documents continue to pose critical research problems.

To address these challenges, various knowledge bases and data models have been developed over time, enabling longitudinal studies, complementary demography, and anthroponymy, which altogether yield deeper insights into population dynamics. The DAI-CRETDHI project specifically focuses on the French population between the 16th and 19th centuries. To capture these dynamics, we employ a dual-track strategy: a macro-level retrospective approach using aggregated data from historical and civil registries, and an "individual approach" applied to select corpora. The latter involves extracting and formalizing a wide array of demographic (sex, age, marital status), family (fertility, household composition, domestic roles), relational (kinship and social networks), socio-professional (occupation, socio-economic status), and geographical (origin, migration status, residence) characteristics. To populate this framework, we rely primarily on data extracted from parish registers and vital records.

The ultimate objective of our project is to utilize digital humanities tools that integrate data modeling and a knowledge base to organize data around the nominal tracking of individual information over medium to long-term periods. This enables not only longitudinal approaches that complement traditional cross-sectional analyses but also observations that extend beyond single points to encompass multiple data points. A key strength of this project is its ability to interconnect with other databases centered on individuals or their units of residence (such as communes or parishes), both existing and future. This will require careful attention to data structuring, standardization, and interoperability. Recognizing named entities (such as people and places) across a vast array of historical documents is essential, yet current extraction and recognition tools struggle with the variability of sources and handwriting styles. Addressing the challenges of geographic coverage and chronological depth will be crucial. Therefore, the application of named entity recognition and spatio-temporal structuring methods is critical to the project, setting it apart with its greater complexity and innovative approach (multi-source and long-term processing). The consortium already possesses some corpora, which are being processed through a combination of participatory data entry and artificial intelligence. Specifically, specialized models for document layout analysis (page segmentation) and handwritten text recognition (HTR) have already been developed and publicly released within the project framework [1]. An example of the document we aim to automatically process is presented in Figure 1. This paper focuses on the foundational data modeling, extraction workflows, and the structuring of the digitized population registers.

2. Scientific context

2.1. Text recognition and Named-Entities management

The performance of Handwritten Text Recognition (HTR) and document layout analysis systems has significantly improved where generated transcriptions are suitable for downstream automated workflows, such as semantic indexing, information extraction, and statistical processing [2, 3]. However, conventional HTR frameworks often remain limited to specific, narrow archival contexts. Historically, these systems required large, homogeneous, and manually annotated datasets from a target corpus [4], usually characterized by a restricted number of scribes or stable administrative layouts [5].

One of the contributions of the DAI-CRETDHI project is to reduce these constraints by processing highly heterogeneous, multi-source parish registers across multiple centuries. Rather than training isolated, closed models, our strategy leverages recent shifts toward flexible, open-science foundations. To this end, specialized models for page segmentation and text recognition have already been trained and

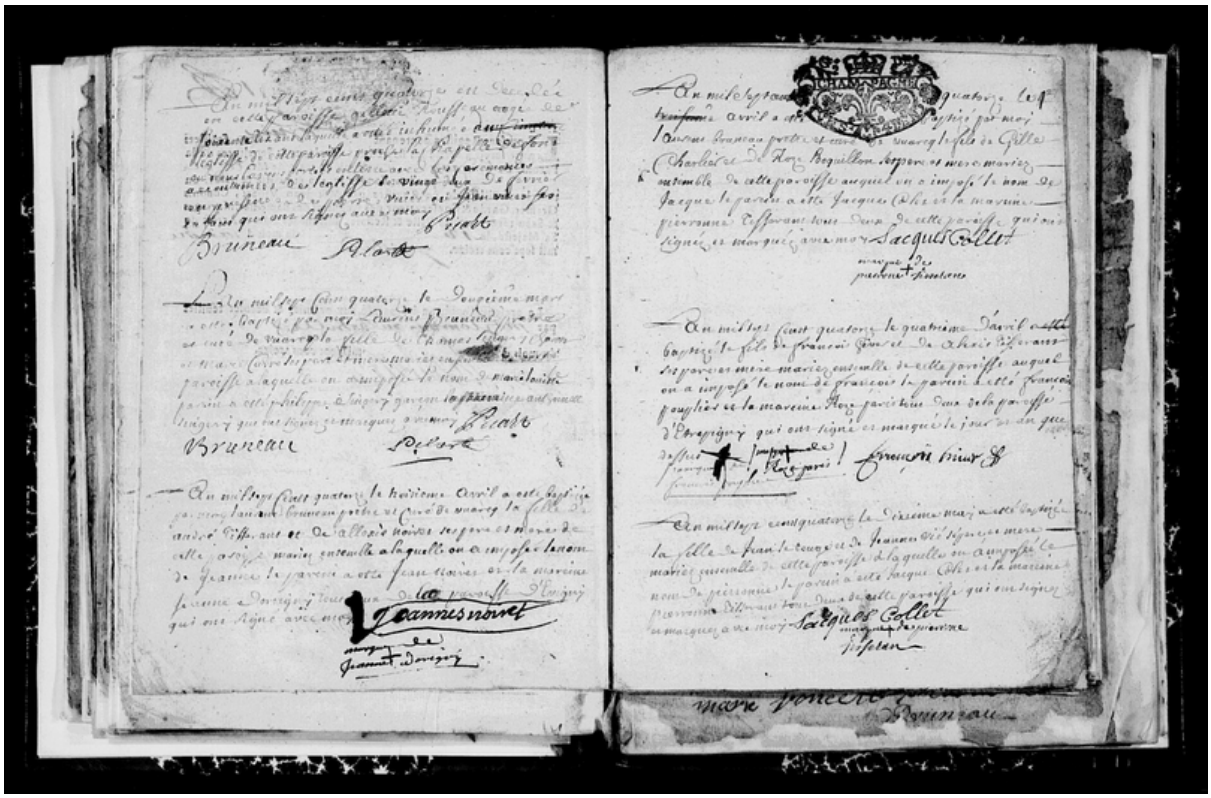


Figure 1: A raw parish register page from the 18th century. Document processing highlights key challenges in historical document analysis, including paper degradation, ink bleed-through from verso pages, and archaic handwriting that requires automated layout extraction, text recognition, and semantic tagging (NER)

publicly released on Hugging Face within the project framework, in collaboration with Teklia [1]. This automated extraction pipeline establishes the groundwork for large-scale, domain-specific Named Entity Recognition (NER), layout analysis, and subsequent entity linking over extensive heritage collections.

2.2. Data management and Knowledge base

The growing availability of digitized historical demographic data allows the consortium to tackle a major historiographical challenge: reconstructing the demographic fabric of France since 1500. This long-standing goal builds upon the pioneer retrospective surveys launched by the French Institute for Demographic Studies (INED) in the late 20th century, which remained unachieved due to technological and scale limitations. Mirroring historical population reconstructions accomplished for England [6] or other national cohorts [7], the objective is to reconstruct long-term population dynamics from aggregated and individual data. While collecting exhaustive records for all 40,000 historical French parishes remains unrealistic, building massive, representative, and interoperable datasets is now a reachable horizon. Recent advances in statistical sampling, big data architectures, and the processing of imperfect or noisy historical data allow us to approach family reconstitution over several generations that was previously difficult to achieve. This large-scale framework provides the empirical corpus necessary to analyze long-term demographic behaviors, kinship systems, and their underlying socioeconomic determinants.

Through this individual-centric approach, DAI-CRETDHI aims to support the restitution of complex life paths (event history analysis) and multi-generational family reconstructions (ascending/descending genealogies, and siblings). Furthermore, nominal tracking allows for the structural analysis of family and social networks, mapping them either as localized ego-networks or as complete, multiplex graphs reflecting diverse layers of historical and domestic relationships.

The representation of such complex, evolving family relationships constitutes a textbook case for

ontology design. For instance, frameworks like the Family History Knowledge Base (FHKB) have formalized intricate kinship structures and rules [8, 9]. While logic-driven approaches and semantic reasoners are highly expressive for deriving implicit genealogical links, they routinely face severe performance and scalability bottlenecks when applied to massive, deep-time datasets. To ensure structural interoperability while maintaining computational efficiency, our architecture relies heavily on established W3C and OGC standards. Temporal dimensions are modeled using the OWL-Time ontology, while spatial and geographic entities are formalized via GeoSPARQL.

The Models and Knowledge team at the L3i laboratory brings extensive experience in deploying these semantic standards across complex spatio-temporal domains. This expertise includes modeling semantic trajectories and mobility patterns.

Given that geospatial anchoring is paramount to tracking individuals across changing administrative boundaries, the DAI-CRETDHI project occupies a strategic, intermediate position between two complementary ANR initiatives. Our ontology serves as a conceptual pivot designed to connect and exploit these distinct historical and geographical data structures for historical demography:

- ANR Obardi (Ontology-based Ancien Régime Data Infrastructure): Focuses on formalizing the institutional, economic, and social structures of 17th and 18th-century France. It provides the Historical Hierarchical Territory (HHT) ontology [10, 11], which models the parish as the fundamental administrative and religious unit of the Ancien Régime.
- ANR Communes: Aims to reconstruct the continuous administrative map of metropolitan France and its transport networks from the 18th century to the present. By integrating the shifting boundaries of all 40,000 communes into a dynamic Geographic Information System (GIS), it allows the temporal tracking of territorial mutations from local to regional levels.

In this ecosystem, the DAI-CRETDHI ontology acts as the necessary intermediary. While historical demographic events occurred within the territorial framework of Ancien Régime parishes, the physical source documents—such as parish registers and vital records—are today preserved by Departmental Archives and legally managed by modern municipal administrations (mapped by ANR Communes). By positioning our knowledge base at the intersection of these two projects, we will deliver a semantic framework capable of relating historical population data to both past institutional contexts and contemporary archival management realities.

3. Project and data

3.1. Historical Background and Consortium

The foundation of the DAI-CRETDHI project builds upon an initial collaborative effort conducted in 2013–2014 between CESR, INED, and the L3i laboratory, titled "*Numérisation, Océrisation, Exploitation: les Enquêtes de Démographie Historique de l'Ined*" (NOE-EDHI). The primary objective of NOE-EDHI was to establish and validate methodological workflows for the retroconversion of nominal data produced by seminal INED surveys (specifically the Henry and Biraben surveys). These retrospective studies were based on the continuous analysis of parish registers across approximately forty sample communes scattered throughout France. Collected over four decades, this massive volume of historical data was originally preserved on paper. Migrating it into structured, accessible digital formats unlocks new opportunities for re-exploiting these surveys and enables novel research, particularly on the oldest nominative records which had never undergone specific computational analysis.

To scale these initial achievements, a new national consortium has been established, combining complementary expertise in historical demography, archival processing, digital humanities, and artificial intelligence:

- **French Institute for Demographic Studies or INED**: A public research institute specialized in population studies and the historical custodian of the baseline demographic surveys.

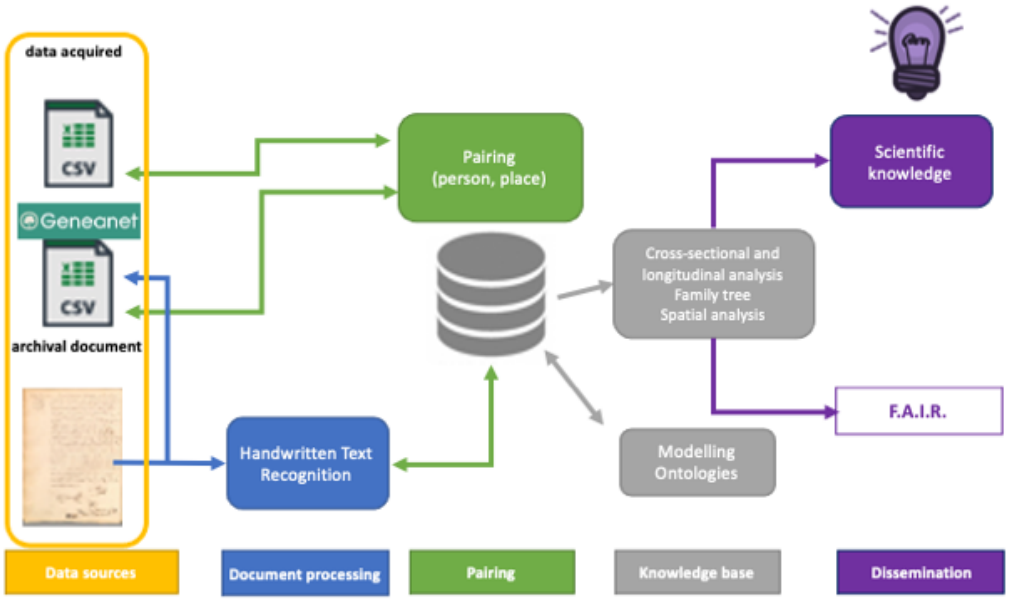


Figure 2: A global view of our project

- **Sorbonne Université - CNRS/ Centre Roland Mousnier (CRM)**: One of the oldest French research centers in history and a foundational pillar of historical demography.
- **Université de Tours / Centre Tourangeau d'Histoire et d'étude des Sources (CeTHiS)**: A research laboratory specialized in editing historical corpora and designing methods for textual and source interpretation.
- **La Rochelle Université / L3i**: The computer science research laboratory of La Rochelle University. Two of its core research teams are actively involved: the Images and Content team, which focuses on document image analysis and layout extraction, and the Models and Knowledge team, which specializes in knowledge engineering, ontologies, and the intelligent management of digital content.
- **Teklia**: A software publisher specializing in artificial intelligence applied to historical documents. Its industrial-grade solutions enable automated layout analysis and handwritten text recognition (HTR) for heritage institutions.
- **Geneanet**: A major genealogical platform hosting a massive collaborative database. It provides access to large-scale, systematically indexed nominative data contributed by an extensive community of users and indexing projects.

3.2. Technical Workflow Architecture

The global architecture proposed by the consortium to process, model, and enrich these historical demographics is summarized in Figure 2. The pipeline is structured around five core functional layers:

- **Data sources**: The input layer consists of heterogeneous historical sources. Researchers and genealogists traditionally rely on tabular spreadsheets for transcription and manual analysis, which are ingested as structured CSV files. Concurrently, the pipeline ingests the corresponding digital images of the original parish register pages.
- **Document processing**: This phase deploys semi-supervised machine learning algorithms to automate information extraction directly from the digitized archival images. The goal is to produce machine-generated transcriptions whose accuracy approaches human-level curation. To support model training, a formal document structure and annotation schema are currently being defined.

- **Pairing**: This intermediate curation step normalizes and structures the raw textual outputs. It focuses on initial entity resolution, syntax cleaning, and the standardized identification of core attributes, such as spatial locations and toponyms.
- **Knowledge base**: Parish and civil registry data are formalized according to their underlying event types (births, marriages, deaths). While source transcriptions vary in formatting across centuries and regions, they share a stable conceptual schema. Semantic mappings are designed to import these transcriptions into the central knowledge base, enabling the system to interconnect distinct historical events associated with the same individual and reconstruct multi-generational family lineages.
- **Dissemination**: The final layer focuses on the open-science distribution of both the trained AI models and the structured datasets. This data infrastructure aims to provide the broader scientific community with a deeper, empirical understanding of long-term population mutations in France.

3.3. Case Study: Cross-Dataset Enrichment

To validate the scalability and integration capabilities of this workflow, the project leverages complementary historical datasets focused on the city of Charleville (located in the Ardennes department).

The legacy 'Charleville' demographic database was historically incomplete, currently containing only individuals whose surnames begin with the letter 'B'. To overcome this limitation, the project utilizes a massive dataset provided by Geneanet, which completed a systematic digital indexing of all available historical archives for the Ardennes department. This external dataset comprises a CSV file containing 4,840,690 rows and 137 columns, where each row represents a demographic event extracted from parish or civil registers.

The core objective is to ingest both the original 'Charleville' database and the Geneanet extraction into the same unified knowledge base. Merging these sources within a single repository introduces the challenge of managing data of variable quality and grain.

By explicitly tracking the origin and lineage of every semantic triplet, the knowledge base allows researchers to query and cross-reference the comprehensive Geneanet data to expand the scope of the 'Charleville' database. Simultaneously, it preserves the ability to filter, weight, or isolate data according to its source and associated reliability level, ensuring robust historical analysis despite the underlying heterogeneity of the corpus.

4. Document and text processing

Historical document analysis and text processing is a mature research domain within the Document Analysis and Recognition (DAR) field, backed by decades of literature published in venues such as the ICDAR conference series [12]. Despite these extensive foundations, processing historical manuscripts remains highly challenging due to substantial variability in handwriting styles, structural layout degradation, and complex named-entity configurations. In our framework, extracting information from parish registers requires solving two consecutive tasks: automated layout analysis (segmenting pages into individual record blocks) and subsequent textual extraction (transcription and semantic tagging).

As illustrated in Figure 1 and Figure 3, historical registers often feature dense, complex tabular or semi-structured layouts where physical separators (lines, columns) are faint, discontinued, or degraded by ink bleed-through. Rather than relying on rigid, rule-based geometric systems that fail across multi-century collections, our pipeline addresses physical structure recognition through an object detection approach.

Specifically, the project deploys a custom architecture derived from the YOLO (You Only Look Once) framework. This model has been specifically adapted and optimized to detect and segment individual records and text paragraphs directly from the raw page inputs. By treatment of layout analysis as an object detection task, the system robustly isolates the boundaries of distinct historical events regardless of local page degradation.

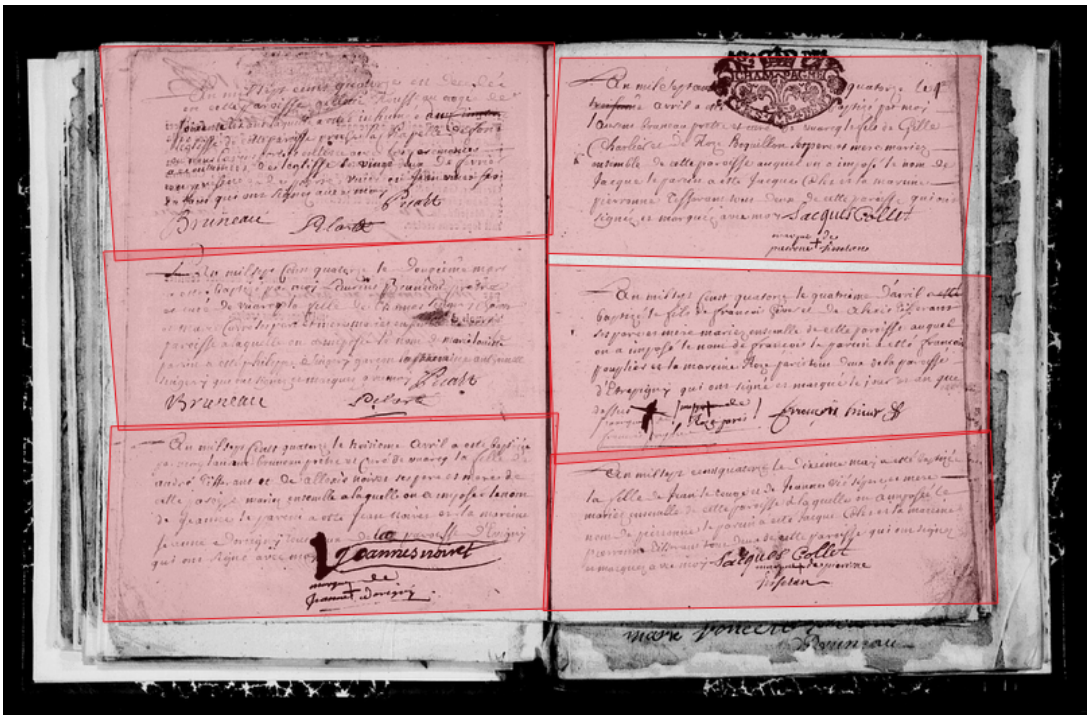


Figure 3: Layout analysis and record segmentation using our custom YOLO-derived model. The red bounding boxes illustrate the automated detection of individual acts and paragraph boundaries on a register page.

Once individual records are segmented, they are passed to the text recognition and information extraction layer. While recent literature heavily investigates end-to-end architectures such as the Document Attention Network (DAN) to generate text and semantic tags simultaneously [13], deploying such models at a national scale requires highly targeted training data. Within the DAI-CRETDHI project framework, a substantial expert annotation campaign has been conducted. Ground-truth data was meticulously annotated to cover both verbatim text transcriptions and the precise anchoring of key named entities (e.g., individuals, toponyms, occupations, and dates).

Currently, the text recognition phase of our production pipeline relies on the reference HTR transcription models developed by Teklia to handle complex character recognition and line-level text generation from handwritten scripts.

While the automated extraction and recognition of named entities (NER) directly from this HTR output represents the next phase of the project, we have initiated our semantic integration workflows using the large-scale data provided by Geneanet. This corpus provides an invaluable foundation because it is already structured around identified entities (such as persons and events). However, given its collaborative nature, it is also highly heterogeneous, presenting challenges of variable cleanliness and a massive density of duplicates.

Utilizing this dataset at this stage allows us to design and stress-test our knowledge base architecture under realistic conditions. By forcing our ontology to ingest and manage voluminous, imperfect entity data, we are developing the necessary semantic alignment and filtering mechanisms. This prepares the infrastructure to robustly handle both the structural redundancy of crowdsourced genealogical data and the native noise that will emerge from the automated HTR pipeline as the project scales.

5. Knowledge database approach

5.1. Ontology Design and Core Concepts

Historical demographers and genealogists routinely rely on tabular spreadsheets to manage records. While data organization varies significantly across individual projects, it remains anchored to the same

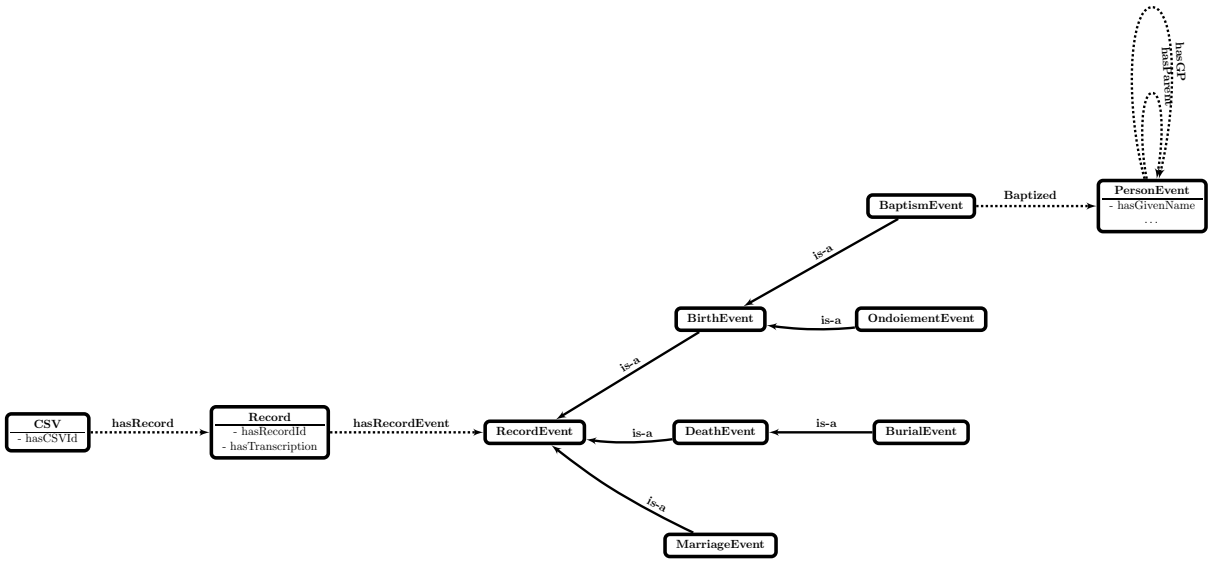


Figure 4: Main concept of our ontology

fundamental domain concepts. Central to the DAI-CRETDHI project is the transition from these flat files to a scalable knowledge base using Semantic Web technologies.

While existing domain frameworks like the Family History Knowledge Base (FHKB) offer predefined mechanisms to model family relationships, they did not fully meet our structural requirements due to the massive volume of data and the computational overhead of logic-driven reasoning over large graphs. Consequently, we designed a custom, streamlined ontology tailored to the specific needs of large-scale historical demography. The core design focuses strictly on direct biological and social ties—specifically parent-child (hasParent) and godparent (hasGP) relationships—while meticulously preserving data provenance back to the original source file. Our model inherits temporal and spatial capabilities by extending the W3C OWL-Time and OGC GeoSPARQL ontologies.

As illustrated in Figure 4, the ontology is structured around four primary conceptual pillars:

- **CSV**: Represents the physical data source (e.g., a specific ingested spreadsheet), capturing metadata like the unique file identifier (hasCSVId) to maintain a continuous chain of data provenance.
- **Record**: Maps individual rows within the source CSV file, where each row corresponds to a specific, dated religious or civil act. It stores contextual metadata (such as row numbers, transcriptions, and source-specific identifiers) and links the document directly back to its parent **CSV** instance.
- **RecordEvent**: Represents the functional and liturgical nature of the recorded act. It structures information specific to the type of sacrament, such as godparents for a baptism or witnesses for a marriage. Our current implementation models six specific historical events:
  - **MarriageEvent**: Captures metadata regarding the bride and groom, their respective parents, and attending witnesses.
  - **BirthEvent**: Serves as a generic concept for civil records or aggregated datasets regarding a newborn. For fine-grained historical sources—particularly prior to the French Revolution—it is specialized into **BaptismEvent** and **OndoiementEvent**.
  - **BaptismEvent**: Specializes *BirthEvent* to model the religious sacrament, capturing the baptized infant, biological parents, birth dates, and the selected godfather and godmother.
  - **OndoiementEvent**: Specializes *BirthEvent* for a simplified, urgent baptism. We intentionally retain the French historical term *ondoiement* as it lacks an exact English equivalent, referring to an emergency baptism in cases of imminent infant death. In historical demography, tracking these events is critical, as high infant mortality rates directly impacted demographic dynamics.

  - **DeathEvent**: Serves as a generic concept for civil records or aggregated datasets regarding a deceased individual. For fine-grained historical sources—particularly prior to the French Revolution—it is specialized into **BurialEvent**.
  - **BurialEvent**: Specializes *DeathEvent* for detailed historical records—especially pre-revolutionary registers—capturing the deceased's profile, the physical burial location, parental ties (for deceased minors), and attending witnesses.
- **PersonEvent**:Structurally decouples individuals from a persistent global identity, capturing instead the specific role, attributes (given names, occupations, age, marital status), and relational context of a person as recorded at the specific moment of the event.

To support transparency and reusability within the semantic web community, the complete OWL source code of our ontology is publicly available at our university repository: https://gitlab.univ-lr.fr/dai-cretdhi-ontology/dai-cretdhi-ontology.

5.2. Triplestore Infrastructure and Scalability

To implement this knowledge base, we adopt a triplestore architecture. This framework allows for the unified management of both our terminological models (TBox) and factual assertions (ABox), streamlining the parsing and querying of highly heterogeneous historical sources.

While a wide range of modern graph databases offer native handling for temporal properties and varying levels of spatial support through GeoSPARQL extensions [22], the final choice of our production triplestore architecture is currently being evaluated. We are actively testing and benchmarking three main environments to determine which setup scales best as our spatial and demographic datasets expand:

- **Apache Jena/Fuseki** and **Ontotext GraphDB**, both widely recognized for their robust compliance with semantic standards and query optimization.
- A custom, forked version of Apache **Marmotta**[3] developed within our team, which introduces specialized spatial geometry enhancements. To ensure environment reproducibility and ease of testing across the consortium, a containerized Docker version of this environment has been made available[4].To meet the high-scalability requirements of processing millions of triples, these architectures are evaluated using a PostgreSQL relational backend empowered by its native PostGIS spatial extension.

The definitive engine selection will be finalized once the large-scale geographical datasets from our partner projects are fully ingested, allowing us to stress-test each triplestore's capacity.

5.3. Entity Linkage and the Duplicate Dilemma

Populating this architecture with massive, multi-source data—including the large-scale Geneanet collections—revealed a fundamental historiographical and technical hurdle. When exploratory SPARQL queries were executed against the triplestore, the results demonstrated a massive density of co-referent entities, redundant records, and duplicate entries.

In historical demography, this redundancy is not merely a data ingestion artifact; it deeply reflects the structural reality of French archival preservation and historical record-keeping, which involves three distinct phenomena:

- **Multi-event individual tracking**: The same real-world individual naturally appears across multiple distinct registers at different points in their life (e.g., at their own baptism, as a witness at a wedding, and finally in a burial registry), often disguised by spelling variations or incomplete metadata.

[3] https://github.com/abouju/marmotta
[4] https://hub.docker.com/repository/docker/abouj/marmotta/

- **Archival duplication (Local vs. *Greffe* collections)**: Historically, French civil and parish registries were legally mandated to be kept in duplicate. For any given event, there exists a local communal register (kept at the parish or municipality) and a *greffe* register (sent annually to the local court/tribunal). Both versions record the same event but are separate physical manuscripts with distinct handwritings, subtle phrasing variations, or differing levels of preservation.
- **Crowdsourcing and Indexing Redundancy (*Tables Décennales*)**: Because the source datasets rely heavily on collaborative indexing, the same event may be transcribed multiple times by independent users. This is further complexified by the inclusion of *Tables Décennales*—alphabetical index tables compiled every ten years since 1792 to summarize births, marriages, and deaths. Ingesting both the original detailed records (from either the local or *greffe* versions) and their corresponding simplified entries in the ten-year indexes inherently creates structural overlaps in the graph.

If left unaddressed, these redundant nodes artificially multiply the observed population size and event counts, heavily distorting downstream demographic indicators, family reconstructions, and longitudinal life-path analyses.

Because resolving these complex, cross-document duplicates directly within a large graph using deterministic rules or native semantic reasoners introduces massive performance bottlenecks, the DAI-CRETDHI project decouples this layer. We utilize Splink, an advanced, open-source probabilistic record linkage framework based on the Fellegi-Sunter methodology. By executing Splink externally (leveraging high-performance execution backends), the project can efficiently compute match probabilities to detect overlapping data entries (resolving the local vs. *greffe* duplicates, or anchoring a *table décennale* entry to its original full record) as well as multi-point individual life paths. Once these probabilistic duplicates are clustered and resolved, the resulting linkage metadata and confidence scores are fed back into the triplestore, allowing our ontology to cleanly isolate, filter, or reconcile co-referent nodes while preserving the demographic integrity of the knowledge base. For instance, regarding the Geneanet dataset for the Ardennes department, the raw CSV file contains 4,840,690 rows. After processing, there are 4,845,386 usable rows, which are reduced to 3,986,632 unique entries once duplicates are eliminated.

6. Conclusion and Perspectives

This paper has introduced the architectural and semantic foundations of the DAI-CRETDHI project (2024-2027), a national infrastructure initiative designed to unlock and model centuries of French historical demographic data. By leveraging state-of-the-art document processing pipelines, we have demonstrated how raw, multi-source parish and civil registers can be transformed into structured digital knowledge. Our current production workflow successfully deploys a custom, YOLO-derived layout analysis model to isolate individual records, which are subsequently transcribed using Teklia's reference HTR models. To ensure reproducibility and foster community collaboration, these specialized models have been publicly released on Hugging Face.

At the core of our approach is a streamlined ontology that extends OWL-Time and GeoSPARQL standards. Rather than enforcing an absolute global entity resolution during the ingestion phase, our model structurally decouples individual historical assertions (PersonEvent) from persistent identities, while rigorously preserving continuous data provenance (CSV and Record metadata). By positioning our knowledge base as a conceptual pivot between the institutional framework of the ANR Obardi project and the geographic continuum of the ANR Communes project, we provide an intermediate framework capable of connecting micro-level demographic events with shifting contemporary and historical territorial jurisdictions.

Several major technical milestones constitute the immediate perspectives of this work. First, our exploration of massive crowdsourced datasets from Geneanet—characterized by real-world noise and intense redundancy—has crystallized a major technical challenge: the proliferation of structural duplicates. In French archives, these duplicates stem not only from multi-event individual life paths but also

from physical archival overlapping (the parallel preservation of local communal registers and court greffe collections) and the inclusion of summarized Tables Décennales.

To resolve these co-referent records without introducing performance bottlenecks within the graph, we are currently integrating Splink, an external probabilistic record linkage framework based on the Fellegi-Sunter methodology. The resulting match probabilities and confidence clusters will be fed back into the graph to reconcile co-referent nodes. Finally, as our geographic and demographic datasets reach national scale, we will finalize our backend infrastructure by benchmarking performance of Apache Jena/Fuseki, Ontotext GraphDB, and our customized version of Apache Marmotta over a robust PostgreSQL/PostGIS foundation. Ultimately, the DAI-CRETDHI infrastructure will deliver a highly scalable, interconnected, and open-access semantic ecosystem for longitudinal studies in French historical demography. Furthermore, as soon as the official archival data from the Departmental Archives (*Archives Départementales*) become fully available, we plan to extend and align our data model with the Records in Contexts-Ontology (RiC-O). Compared to event-centric museum ontologies such as CIDOC-CRM, RiC-O appears more naturally suited to our corpus, as it specifically formalizes archival structures, document provenance, and record-keeping practices. This alignment will ensure seamless interoperability with broader archival descriptions and standard institutional metadata structures.

6.1. Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors used [Gemini, Claude, ChatGPT] for two main purposes: (1) to verify and improve the grammatical accuracy, readability, and stylistic flow of the manuscript, and (2) to cross-verify the structural correctness, syntax, and alignment of the data models, ontologies, and technical pipelines described herein. Following this process, the authors thoroughly reviewed, tested, and validated all textual and conceptual outputs, and take full responsibility for the final architecture, integrity, and overall content of the published paper.

Acknowledgments

This research work has been funded, in full or in part, by the French National Research Agency (ANR) under the agreement "ANR-23-CE38-0016-01". We would like to thank François-Joseph Ruggiu, Vincent Gourdon and Isabelle Robin from CRM, Antoine Doucet from L3i, Jérôme Galichon and Aliénor Samuel-Hervé from Geneanet, Christopher Kermorvant and Solène Tarride from TEKLIA for their contributions to this paper.

References

[1] DAI-CRéTDHI Project Model Collection, https://huggingface.co/collections/Teklia/dai-cretdhi-project, 2025. Hugging Face Repository, Accessed: June 2026.

[2] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, A. Doucet, N. Sidere, M. Coustaty, Melhissa: a multilingual entity linking architecture for historical press articles, International Journal on Digital Libraries 23 (2022) 133–160. URL: https://doi.org/10.1007/s00799-021-00319-6. doi:10.1007/s00799-021-00319-6.

[3] H. Vézina, C. Kermorvant, M. Bonhomme, J. Bournival, i-balsac: Completing families with the help of automatic text recognition, Social Science History Association (2019) 1–25.

[4] G. Muehlberger, L. Seaward, M. Terras, S. A. Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, et al., Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, Journal of documentation 75 (2019) 954–976.

[5] T. Constum, N. Kempf, T. Paquet, P. Tranouez, C. Chatelain, S. Brée, F. Merveille, Recognition and information extraction in historical handwritten tables: Toward understanding early 20 th century paris census, in: International Workshop on Document Analysis Systems, Springer, 2022, pp. 143–157.

[6] E. A. Wrigley, R. S. Schofield, The population history of England 1541-1871: A Reconstruction, London: Edward Arnold, 1981.
[7] M. Voutilainen, J. Helske, H. Högmander, A bayesian reconstruction of a historical population in finland, 1647–1850, Demography 57 (2020) 1171–1192.
[8] R. Stevens, M. Stevens, N. Matentzoglu, S. Jupp, Manchester Family History Advanced OWL Tutorial, 1.0 ed., University of Manchester, 2013. URL: http://owl.cs.manchester.ac.uk/tutorials/fhkbtutorial/.
[9] S. T. Robert Stevens, Nicolas Matentzoglu, Family history knowledge base (fhkb) tutorial v2.0, 2015. URL: https://oboacademy.github.io/obook/tutorial/fhkb/.
[10] W. Charles, N. Aussenac-Gilles, N. Hernandez, Hht: An approach for representing temporally-evolving historical territories, in: C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2023, pp. 419–435.
[11] W. Charles, N. J. Hernandez, HHT : An ontology to represent territorial dynamics for digital humanities, in: N. Aussenac-Gilles, A. Glaton, T. Hamman, M. M. Hedblom (Eds.), 13th International Conference on Formal Ontolog in Information Systems (FOIS 2023), volume 377 of *Frontiers in Artificial Intelligence and Applications*, Nathalie Aussenac-Gilles and Antony Galton and Torsten Hamman and Maria M. Hedblom, IOS Press, Sherbrooke, Canada, 2023, pp. 257 – 271. URL: https://hal.science/hal-04310997. doi:10.3233/FAIA231133.
[12] E. H. B. Smith, M. Liwicki, L. Peng (Eds.), Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, GREECE, August 30–September 4, 2024, Proceedings, Part I, volume 14807 of *Lecture Notes in Computer Science*, Springer, 2024.
[13] D. Coquenet, C. Chatelain, T. Paquet, Dan: A segmentation-free document attention network for handwritten document recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 8227–8243. doi:10.1109/TPAMI.2023.3235826.