Mapping Change: A Temporal and Semantic Knowledge Base of Scottish Gazetteers (1803–1901)

Lilin Yu, **Rosa Filgueira**
EPCC, University of Edinburgh
L.Yu-40@sms.ed.ac.uk, r.filgueira@epcc.ed.ac.uk
Identifier: https://rosafilgueira.github.io/MappingChange-Paper-ISWC2025/

Abstract

We present **MappingChange**, a project that constructs a temporal and semantic knowledge base from ten 19th-century **Gazetteers of Scotland** (1803–1901), digitized as over 13,000 page-level **XML** files. These noisy, unstructured texts lack article-level markup and exhibit highly heterogeneous layouts. To segment and structure over 50,000 historical place descriptions, we employ **large language models (LLMs)** with **edition-specific prompting strategies**, tuned to handle distinct editorial conventions, abbreviations, and multi-page entries. The resulting knowledge base comprises three interlinked knowledge graphs: (1) a **basic KG**, extracted from cleaned DataFrames; (2) a **concept-enriched KG**, linking semantically similar place records across editions using **sentence embeddings**, **Wikidata**, and **DBpedia**; and (3) a **location-annotated KG**, enriched with **named entity recognition** and **geographic disambiguation**. We further align extracted entities to external sources such as Wikidata and DBpedia, enabling rich contextual integration and reuse. All are expressed in **RDF** and modeled with the updated **Heritage Textual Ontology (HTO)**, which provides a structured vocabulary for capturing textual provenance, bibliographic metadata, extraction context, and diachronic semantic alignment across editions. In addition to the knowledge graphs, we release: (a) individual DataFrames for each edition, (b) a unified cross-edition DataFrame, and (c) Elasticsearch indices. All resources are integrated into the Frances(http://www.frances-ai.com) semantic web platform, enabling historical exploration through keyword and semantic search, concep-timeline navigation, and interactive geolocation visualizations.

1. Introduction

Descriptive gazetteers were central to how 19th-century Scotland documented its geography—towns, parishes, rivers, castles, lochs, and glens—embedding each place within broader historical, economic, and social narratives. These texts evolved over the century, reflecting transformations brought about by industrialization, land reform, public health, and imperial expansion. The *Gazetteers of Scotland, 1803–1901*, digitized by the National Library of Scotland (NLS), constitute one of the most comprehensive corpora for studying Scotland's spatial knowledge in the long 19th century. The full collection comprises twenty

volumes produced by different publishers and editors, and has been released as more than 13,000 high-resolution scans accompanied by ALTO XML files. These XML files encode layout and textual content extracted via Optical Character Recognition (OCR), resulting in over 1.75 million lines and 14 million words. While this makes the data technically accessible, it remains largely unsuitable for structured analysis: the texts lack article-level markup, exhibit inconsistent typographic structures, and contain significant OCR noise. Entries often begin mid-column, span multiple pages, and vary widely in format and editorial style—posing major challenges for computational processing, information retrieval, and historical reuse.

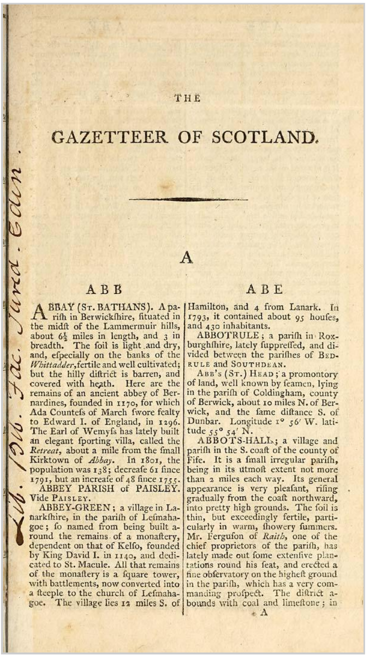
Compounding these challenges is the fact that many place names (e.g., "ABBEY" or "GREENHILL") recur across the gazetteers, often referring to different locations. Disambiguating such entries is non-trivial, as it depends on contextual clues within each article rather than surface-level patterns. Our approach relies on LLM-based article segmentation and interpretation—capturing subtle editorial cues and semantic context to correctly associate each name with the appropriate description.

MappingChange is the first project to construct a structured, queryable, and semantically enriched temporal knowledge base from this entire collection. We extract and align over 50,000 historical place descriptions across ten gazetteer editions, using large language models (LLMs) and volume-specific prompting strategies that are carefully tuned to editorial idiosyncrasies. The result is a knowledge base composed of three interlinked knowledge graphs: a basic graph derived from structured DataFrames; a concept-enriched graph linking semantically similar entries across editions using sentence embeddings, Wikidata, and DBpedia; and a location-annotated graph generated through geographic disambiguation techniques. These graphs are serialized in RDF and modeled using the Heritage Textual Ontology (HTO), a domain ontology we developed specifically for historical and heritage corpora.

The **Heritage Textual Ontology (HTO)** (see resource here) is designed to model not just entities and attributes but also the editorial and computational processes by which each record is extracted, cleaned, and enriched. Unlike generic ontologies, HTO supports the representation of textual provenance, extraction prompts, editorial hierarchies, and diachronic linkage across editions. It enables us to track how descriptions of the same place evolve over time, with full transparency into their source structure and transformation process. Its design has been guided by real-world use cases in digital heritage, and it plays a central role in making the resulting knowledge graphs both expressive and reproducible.

The complexity and variability of these sources can be seen in Figure 1, which presents the opening pages of two editions: the 1803 *Gazetteer of Scotland* and the 1884 *Ordnance Gazetteer of Scotland*. These differences, compounded across volumes, necessitate a custom approach to segmentation, prompting, and post-processing—especially since no

edition includes machine-readable metadata or reliable article delimiters. Note that articles can span multiple pages, but there is no page-level indication of article continuation.



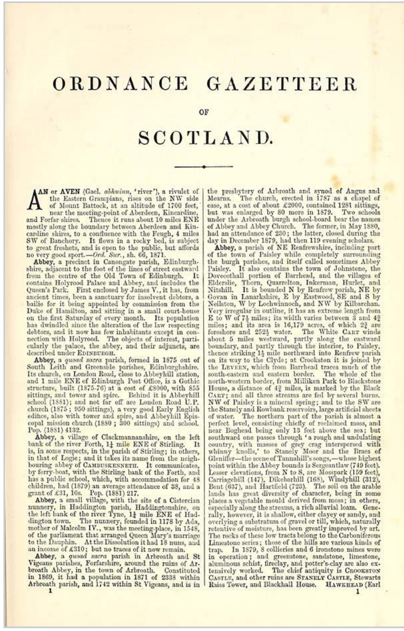
THE

GAZETTEER OF SCOTLAND.

A

ABB ABE

Figure 1 (left): Opening entries of the 1803 *Gazetteer of Scotland*. Page headers consist of two three-letter uppercase segments (e.g., “ABB ABE”). Place names appear in all caps, typically followed by a period or semicolon—offering minimal typographic separation between entries.



ORDNANCE GAZETTEER

OF

SCOTLAND.

Figure 1 (right): Opening entries of the 1884 *Ordnance Gazetteer of Scotland*. This edition presents visually clearer structure, with entries formatted in title case and followed by commas. However, it introduces long, multi-paragraph articles with historical digressions and cross-references—necessitating more refined prompting strategies for accurate segmentation.

Of the twenty volumes in the NLS dataset, we process ten as fully descriptive gazetteers with complete metadata and dual-volume structure. We exclude the 1828 edition, which is a town-focused summary rather than a gazetteer, and the 1848 edition, for which only Volume II survives. All ten processed volumes are mapped into structured DataFrames, knowledge graphs, and Elasticsearch indices that enable full-text and vector-based semantic search. These resources are deployed via the Frances semantic web platform, allowing users to explore the evolution of Scottish place descriptions through SPARQL queries, concept timelines, and geolinguistic visualizations. All code and data are openly available at github.com/francesNLP/MappingChange.

This work offers a reusable digital resource that transforms a historically important but structurally inaccessible corpus into a machine-readable knowledge base for temporal, geographical, and linguistic analysis—paving the way for new forms of linked data research in cultural heritage and historical geography.

The remainder of this paper is structured as follows. Section 2 reviews related work on gazetteer digitization, knowledge base construction, and the use of language models for

historical document processing. Section 3 introduces the Heritage Textual Ontology (HTO), including its conceptual model and alignment with other semantic vocabularies. Section 4 details the end-to-end *MappingChange* pipeline, from OCR ingestion and LLM-based segmentation to DataFrame creation and RDF serialization. Section 5 presents the construction of the three interlinked knowledge graphs and their semantic enrichment through embeddings, entity linking, and location annotation. Section 6 describes the integration of all outputs into the Frances semantic platform, highlighting search, querying, and visual exploration capabilities. Section 7 provides qualitative examples and a usage scenario illustrating how the knowledge base supports temporal and comparative analysis of Scottish place descriptions. Section 8 concludes with a summary of contributions and discusses directions for future work.

2. Related Work

Efforts to structure historical textual data have increasingly turned to Semantic Web technologies. Projects like Linked Places, WarSampo, and Enslaved.org demonstrate how knowledge graphs can model entities, events, and relationships from historical sources. Similarly, national libraries have applied OCR and metadata modeling to digitize collections, such as the NLS's Data Foundry.

Our work builds on this tradition while addressing unique challenges presented by 19th-century gazetteers: noisy OCR, lack of structural markup, and inconsistent editorial conventions. Prior methods using rule-based extraction or statistical models fall short when applied to these irregular texts. Recent advances using large language models (LLMs) like GPT-4 show promise in segmenting, interpreting, and linking historical content. We apply such models at scale, with edition-specific prompting strategies and custom cleaning heuristics, to generate reliable article-level entries and semantic annotations.

Our ontology-driven approach aligns with recent efforts like the EB Ontology, NLS Ontology, and the Enslaved Ontology, but introduces the more flexible and extensible Heritage Textual Ontology (HTO), designed specifically for long-range temporal comparison and provenance tracking across multiple editions and digitization methods.

3. Resource Description

Mapping Change includes:

- Ten digitized Gazetteers of Scotland (1803–1901) from the NLS Digital Repository.
- Over 50,000 extracted articles, segmented and structured using GPT-4 prompts tailored to each volume.

- Volume-specific JSON DataFrames enriched with metadata, identifiers, and embeddings.
- An RDF/Turtle knowledge graph using the HTO ontology.
- Entity links to Wikidata and DBpedia.
- Elasticsearch indices supporting both keyword and vector similarity search.
- Full integration with the Frances semantic web platform.

All code and data are publicly available via github.com/francesNLP/MappingChange.

4. Heritage Textual Ontology

The Heritage Textual Ontology (HTO) provides the semantic backbone for *Mapping Change*, enabling structured representation of historical textual records, their provenance, and evolving place-based concepts. Since its initial release, HTO has undergone substantial refinement to support richer semantic modeling, improved interoperability, and enhanced tracking of digitization workflows and AI-assisted outputs.

The ontology is openly developed at github.com/frances-ai/HeritageTextOntology, and its documentation, including diagrams and examples, is available at w3id.org/hto.

4.1 Key Ontological Enhancements

- **Textual Record Modeling**: New classes such as `HTO:Article`, `HTO:PlaceRecord`, `HTO:InternalRecord`, and `HTO:TermRecord` differentiate between OCR-extracted fragments, cleaned entries, and semantically disambiguated concepts. The `HTO:Description` class tracks structured outputs from GPT-4, manual annotations, or post-processing tools.

- **Digitization Provenance**: Bibliographic metadata is modeled using `HTO:Work`, `HTO:Volume`, and `HTO:Edition`, with provenance relationships defined via PROV-O and schema.org. Each `HTO:Article` is linked to its digitized source via permanent NLS page URLs and includes annotations such as `HTO:textQuality` to assess OCR accuracy and reliability.

- **LLM Provenance and Prompt Modeling**: The new class `HTO:InformationResource` represents prompt templates, LLM configurations, and model-generated outputs, enabling full traceability of GPT-based extractions. This supports transparent reuse, auditing, and future replication.

- **Concept Evolution and Semantic Clustering**: `HTO:Concept` is used in combination with SKOS to group equivalent or evolving place references across multiple gazetteer editions. Concepts can represent locations, institutions, or

geographical types and are dynamically inferred from embeddings and term clustering.

- **Geographic and Type Annotation**: The ontology introduces `HTO:GeographicAnnotation` for storing lat/lon coordinates derived from external services or contextual inference. It also includes `HTO:LocationType` for classifying place categories (e.g., parish, river, estate).
- **Linking and External Alignment**: Instances of `HTO:PlaceRecord` and `HTO:Concept` may include links to external resources using `HTO:externalMatch`, allowing interconnection with Wikidata, DBpedia, and other knowledge bases.
- **Lineage and Versioning Support**: Using `HTO:wasDerivedFrom`, `HTO:wasRecordedIn`, and `HTO:hasTextQuality`, the ontology supports full lineage tracking from OCR to human-reviewed RDF. This is critical for understanding transformations across stages of digitization, modeling, and enrichment.

4.1 Example Use in Mapping Change

Each gazetteer article is instantiated as an `HTO:Article`, linked to its originating `HTO:Volume` and to one or more `HTO:Concept`s (e.g., “Aberdeen”). Concepts aggregate variations of place descriptions across editions, while RDF-level annotations record when, where, and how each article was extracted or transformed.

Prompt templates and GPT outputs are represented as `HTO:InformationResource`s, allowing clear documentation of AI-assisted steps. This structured metadata facilitates reproducibility and comparative studies across digitized corpora.

HTO is designed to be extensible and aligns with best practices in cultural heritage modeling, combining traditional bibliographic ontologies with novel AI-aware components.

5. Construction and Content

The resource was built using a modular pipeline comprising:

5.1 Extraction

- Volume-specific scripts (e.g., extract_gaz_1803.py) segment OCR text using GPT-4 with prompts adapted to differing article structures.
- Prompts handle varying formats, including mid-page entries, redirects, and irregular headers.

Volume-Specific Prompt Engineering

Because each Gazetteer edition between 1803 and 1901 features highly distinct layout conventions (e.g., capitalization, abbreviations, header formatting, article delimiters), we could not apply a single uniform prompt across all volumes. Instead, we designed **custom GPT-4 prompts** for each edition to ensure accurate article segmentation and place name extraction.

The table below summarizes the key differences and our adaptation strategies:

Gazetteer Volume	Prompt Focus	Format Characteristics	Prompt Adaptation Strategy
1803	Entry detection in irregular formatting	Short entries, inconsistent punctuation	Prompt includes examples with minimal structure; stresses sentence-level cues for boundaries
1806	Parsing longer headers	Descriptive headers like "Parish of…"	Prompt highlights multi-word headers and requests exact header extraction
1825	Delimiting fused articles	Minimal line breaks between articles	Prompt stresses lexical patterns (e.g., place types, initial caps) to find boundaries
1838	Handling abbreviations and symbols	Use of brackets, abbreviations for counties	Prompt includes example abbreviations and instructions to include them in headers
1842	Identifying hierarchical entries	Entries with sub-places or parenthesized detail	Prompt uses hierarchical examples and specifies nested JSON structure
1846	Normalizing inconsistent capitalization	Random capital words mid-paragraph	Prompt emphasizes ignoring internal caps unless followed by specific patterns
1868	Filtering out printed annotations	Use of special characters, side notes	Prompt includes rule to ignore marginal notes or typesetting artifacts
1884 & 1901	Unified structured prompt	Consistent bold headers, clear formatting	A single prompt applied to both; relies on standard visual patterns and separators

Each prompt is represented as an instance of `HTO:InformationResource`, enabling traceable documentation of prompt design and LLM usage in our pipeline.

5.1 Cleaning & Deduplication

- Cleaned JSON outputs are merged.
- Fuzzy matching, prefix-trees, and substring containment detect duplicates across years and within volumes.

5.1 DataFrame Generation

- Unified metadata from OCR, XML, and GPT outputs are exported to structured JSON-based DataFrames.

5.1 Knowledge Graph Generation

- RDF triples are created using the improved HTO ontology.
- Entities include Articles, Volumes, Concepts, and digitization provenance.

5.1 Entity Linking

- Gazetteer terms are matched to DBpedia and Wikidata using label and description matching.
- Articles with similar embeddings are grouped into concepts using `all-mpnet-base-v2`.

5.1 Enrichment

- Concepts are assigned summaries, sentiment values, and external links.
- Article timelines visualize the evolution of place concepts across editions.

5.1 Search Indices

- Elasticsearch indices are built for articles, Wikidata, and DBpedia entities.
- Vector search enables semantically similar article discovery.

5.1 Geoparsing

- For enriched geospatial analysis, `geoparse.py` tags locations using SpaCy NER and Gazetteer context.

All scripts are in MappingChange/src/, and their outputs are versioned and archived.

6. Usage

Mapping Change can be explored in three main ways:

6.1 Data Access

- All cleaned DataFrames and RDF graphs are in the GitHub repository: francesNLP/MappingChange
- Scripts for reproducing those dataframes, KGs and ES are in the GitHub repository: francesNLP/MappingChange
- Zenodo DOI (to be added)

6.1 SPARQL Querying

- A Fuseki SPARQL server supports knowledge graph exploration.
- Sample queries for retrieving places, concepts, and links are included.

6.1 Frances Platform

- Users can search and explore articles via full-text or semantic search.
- Concepts are visualized through timelines and embeddings.

6.1 Notebooks

Google Colab notebooks are provided for each gazetteer to enable direct exploration and analysis.

7. Sustainability

Mapping Change is designed to support long-term historical research:

- **Archiving**: All code, data, and RDF outputs are versioned and archived on Zenodo.
- **Ontological Reuse**: The HTO ontology is maintained and extended in an open repository with permanent identifiers.
- **Frances Platform Integration**: The data is accessible through a production-ready semantic platform, ensuring ongoing usability beyond the scope of the project.
- **Extensibility**: The pipeline is modular and supports the integration of new volumes, editions, or other regional gazetteers.

8. Conclusion

Mapping Change creates a temporal, semantic infrastructure for exploring Scottish place descriptions from 1803–1901. Combining LLM-based extraction, improved ontology design, and semantic search, we deliver a reusable, interoperable dataset for historical research.

The improved HTO ontology enables robust modeling of textual provenance, record quality, and evolving concepts. The Frances platform empowers researchers to query and visualize this data across time and space.

Future work includes integrating cartographic metadata, and link it to the 100 years of the Encyclopaedia Britannica.

9. Acknowledgements

This work was supported by the Royal Society of Edinburgh (RSE Small Research Grant).

@article{semanticweb, title = {The semantic web}, author = {Berners-Lee, Tim and Hendler, James and Lassila, Ora and others}, journal = {Scientific American}, volume = {284}, number = {5}, pages = {28–37}, year = {2001}, publisher = {New York, NY, USA:} }