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, place-name disambiguation, 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**. 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 semantic web platform, enabling historical exploration through keyword and semantic search, as well as through interactive visualizations.

1. Introduction

Descriptive gazetteers were a cornerstone of how 19th-century Scotland recorded and transmitted its geographical knowledge—capturing towns, parishes, rivers, castles, and natural features within evolving historical, social, and economic narratives. As industrialization, migration, land reform, and empire reshaped the country, these texts became critical instruments for documenting change. The *Gazetteers of Scotland, 1803–1901*, digitized by the National Library of Scotland (NLS), form one of the most extensive corpora for studying this spatial transformation. Yet despite being released as more than 13,000 OCR-aligned ALTO XML files, the collection remains largely unsuitable for structured analysis. The lack of article-level markup, noisy layouts, and inconsistent editorial conventions limit its accessibility for digital scholarship.

Each edition introduces distinct typographic conventions for article headers, redirects, abbreviations, and multi-page entries. In early volumes, place names appear in uppercase with minimal punctuation to separate entries (Figure 1, left), whereas later editions use clearer formatting, including title casing and consistent delimiters (Figure 1, right). These variations complicate traditional rule-based approaches to text segmentation and alignment, making it difficult to trace how a place is described across time or to resolve recurring names that refer to different locations.

Compounding these challenges is the fact that many place names (e.g., "ABBEY" or "GREENHILL") recur across the gazetteers, often referring to different locations. Moreover, later editions tend to include a broader set of places, meaning some names appear for the first time in later volumes or gain more detailed descriptions over time. 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.

To overcome these limitations, we present **MappingChange**, a reusable and openly licensed resource that transforms this historically rich but structurally fragmented corpus into a structured and semantically enriched knowledge base. Our pipeline uses large language models (LLMs) with edition-specific prompt strategies to extract over 50,000 article-level entries and express them in RDF as three interlinked knowledge graphs, modeled using the Heritage Textual Ontology (HTO). The resulting knowledge base supports both **temporal analysis**, enabling comparison of how place descriptions change across editions, and **semantic exploration**, through the linking of conceptually related entries and connections to Wikidata and DBpedia.

A full account of the dataset structure, file formats, and access methods is provided in Section 3. All components—including JSON DataFrames, RDF knowledge graphs, and Elasticsearch indices—are made available via github.com/francesNLP/MappingChange and integrated into the Frances semantic platform, supporting FAIR access, reproducibility, and visual exploration.

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.

Figure 1 (right): Opening entries of the 1884 *Ordnance Gazetteer of Scotland*. This edition features a clearer visual structure, with entries formatted in title case and followed by commas. Page headers display the first and last place names on the page, both rendered in uppercase.

By providing not only a semantically rich knowledge base but also a fully documented and modular pipeline for its construction, this work contributes to the broader goals of the Semantic Web and digital heritage. All components—DataFrames, RDF knowledge graphs, and Elasticsearch indices—are reproducible from source using openly available scripts, which are detailed in Section 5. This enables other researchers to reuse, adapt, or extend the MappingChange infrastructure for new corpora, ensuring FAIR data practices and long-term sustainability. In doing so, the project facilitates new forms of historical geography and cultural analytics by making a previously inaccessible corpus computationally interoperable, queryable, and reusable across domains.

The remainder of this paper is structured as follows. Section 2 reviews related work on gazetteer digitization, knowledge base construction, and the use of large language models for historical document processing. Section 3 provides a detailed description of the MappingChange resource, its components, and access mechanisms. Section 4 details the Heritage Textual Ontology (HTO), including its conceptual model and alignment with existing semantic vocabularies. Section 5 outlines the end-to-end pipeline for extraction, cleaning, and semantic enrichment. Section 6 presents usage scenarios and queries that demonstrate how the resource supports historical research. Finally, Section 7 concludes with a summary of contributions and future directions.

2. Related Work

A growing body of research in digital humanities and cultural heritage has increasingly adopted Semantic Web technologies to structure, enrich, and interlink historical textual corpora. Notable examples include WarSampo [1], which models World War II data from Finland as Linked Open Data, and Enslaved.org, which applies graph-based modeling to records from the transatlantic slave trade. These projects demonstrate how knowledge graphs can be used to represent complex relationships among people, places, and events in heterogeneous historical sources.

At the national level, initiatives such as the Europeana Data Model EDM [2] and the National Library of Scotland's Data Foundry exemplify large-scale digitization and metadata modeling efforts aimed at improving accessibility and reuse of cultural heritage data. Europeana promotes interoperability through linked data principles and vocabulary standardization, while the NLS provides high-quality scans and ALTO XML for thousands of 19th-century documents, including the Gazetteers of Scotland.

Nevertheless, these infrastructures alone are insufficient for corpora like Scottish gazetteers, which present significant challenges: noisy OCR, lack of article-level segmentation, mid-page article starts, and inconsistent editorial conventions across editions. Traditional approaches to structuring such texts—including rule-based or statistical methods—often fail under these conditions. Prior work on historical textual collections, such as newspapers or the Encyclopaedia Britannica (e.g., using the `defoe` library [3] has demonstrated the need for scalable, domain-adapted pipelines.

Recent breakthroughs in large language models (LLMs) such as GPT-4 [4] open new possibilities for flexible text interpretation and segmentation. Our project leverages these models at scale with custom prompts tailored to the editorial style of each edition. This enables us to segment and extract over 50,000 structured article-level entries from ten 19th-century gazetteer volumes, while handling abbreviation styles, redirects, and evolving toponym usage.

Beyond extraction, semantic modeling is critical to ensuring data reusability and interpretability. We build on and extend prior ontologies developed for cultural heritage contexts—such as the Encyclopaedia Britannica Ontology [5] and the National Library of Scotland Ontology [6]—to model bibliographic provenance and source structure. However, these earlier ontologies were not designed to represent the full diachronic and computational transformation history of digitized corpora.

To address this gap, we introduced in our previous work [7] the Heritage Textual Ontology (HTO [8]), which provides a provenance-aware semantic framework that models not only the source structure and bibliographic metadata, but also digitization context, Name Entity Recognition (NER)-based outputs, and semantic enrichments. HTO integrates concepts from PROV-O [9] and Schema.org, but extends them with domain-specific classes and

properties tailored to heritage corpora, supporting integration of various data sources from the same corpora.

Finally, our work is fully integrated into the Frances [7] semantic platform, which supports temporal exploration and semantic search of historical data. Recent updates to Frances, include improved support for concept clustering, knowledge graph visualization, and extensible RDF modeling, making MappingChange a robust and reusable infrastructure for temporal knowledge base construction in historical research.

3. Resource Description

The MappingChange resource provides a reusable, modular knowledge base that transforms the digitized Gazetteers of Scotland (1803–1901) into a structured and semantically enriched dataset for historical place-based analysis. Central to the resource are volume-specific JSON DataFrames containing over 50,000 articles—each corresponding to a place entry—extracted from OCR-aligned ALTO XML using edition-specific GPT-4 prompts. These DataFrames include article text, identified place names, and structured metadata such as edition, volume, page number, and word count. They serve as the foundation for the construction of three interlinked RDF knowledge graphs, modeled using the Heritage Textual Ontology (HTO), which captures both textual provenance and semantic transformations.

The first knowledge graph encodes cleaned article-level records and their bibliographic context, preserving references to their original digitized pages. The second graph introduces concept-level alignment by clustering semantically equivalent entries across editions, assigning persistent identifiers and linking them to external entities such as Wikidata and DBpedia. The third graph adds geographic enrichment, using named entity recognition and georesolution techniques to annotate articles with location coordinates and spatial types. Together, these layers enable advanced temporal and semantic analysis of how places were described over time in the Scottish Gazetteers.

All knowledge graphs are serialized in RDF/Turtle and adhere to FAIR data principles. Outputs are hosted in a public Fuseki SPARQL endpoint and are complemented by a set of Elasticsearch indices that support both traditional keyword search and vector-based semantic similarity queries. These indices cover articles, concepts, and external entity links, enabling fast and flexible retrieval across editions and enrichment layers. All components are deployed within the Frances semantic platform, providing users with interactive interfaces for querying, timeline visualization, and concept exploration.

The full resource—including DataFrames, RDF files, and search indices—is openly available at https://github.com/francesNLP/MappingChange, with persistent identifiers to be issued via Zenodo. The HTO ontology is developed in an open repository under a CC-BY 4.0 license. To promote reuse and reproducibility, the repository includes modular

scripts and comprehensive documentation. The step-by-step process by which these resources are constructed—including article extraction, prompt design, semantic clustering, and enrichment—is detailed in Section 5.

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. It 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 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.

- **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. 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.

8. Acknowledgements