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 resource that constructs a temporal and semantic knowledge base from ten 19th-century **Gazetteers of Scotland** (1803–1901), digitized as over 13,000 page-level **ALTO XML** files. These texts are noisy, inconsistently structured, and lack article-level markup—posing substantial challenges for reuse. To segment and extract over 50,000 historical place descriptions, we employ **large language models (LLMs)** with **edition-specific prompt strategies**, tuned to handle diverse editorial conventions, abbreviations, recurring place names, and multi-page entries. The resulting knowledge base comprises three interlinked knowledge graphs: (1) a **basic KG**, representing extracted article-level entries and their metadata; (2) a **concept-enriched KG**, clustering semantically related descriptions across editions using **sentence embeddings**, with links to **Wikidata** and **DBpedia**; and (3) a **location-annotated KG**, enriched through **named entity recognition** and **georesolution**. All graphs are expressed in **RDF** and modeled using the extended **Heritage Textual Ontology (HTO)**, which captures textual provenance, extraction context, and diachronic alignment. In addition to the knowledge graphs, we release (a) edition-specific and aggregated DataFrames (one row per place entry, including metadata and embeddings), and (b) a set of Jupyter Notebooks illustrating usage of our resource. All components are openly available and deployed through the Frances semantic platform and a public Fuseki endpoint, enabling historical exploration via semantic search, timeline comparison, and visual querying interfaces.

1. Introduction

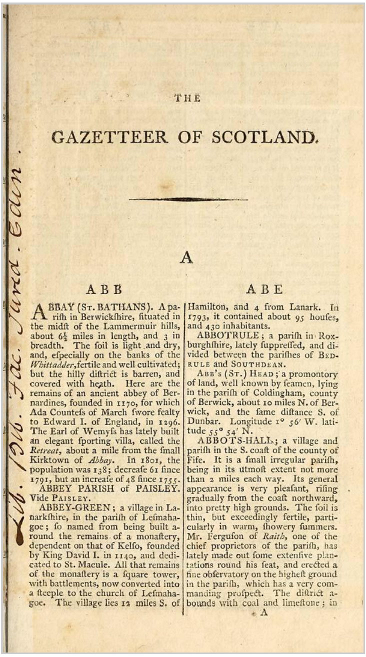
Descriptive gazetteers were a cornerstone of how 19th-century Scotland recorded and transmitted its geographical knowledge—capturing towns, parishes, rivers, castles, and natural features within evolving historical, social, and economic narratives. As industrialization, migration, land reform, and empire 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") 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 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 resources (DataFrames, knowledge graphs, and Notebooks) are publicly archived and distributed via our repository, with persistent identifiers issued via Zenodo and integrated into the Frances semantic web platform for visual exploration.


THE

GAZETTEER OF SCOTLAND.

A

ABB ABE

ABBAY (St. BATHANS). A pariſh in Berwickſhire, ſituated in the midſt of the Lammermuir hills, about 6½ miles in length, and 3 in breadth. The ſoil is light and dry, and, eſpecially on the banks of the *Whittadder*, fertile and well cultivated; but the hilly diſtrict is barren, and covered with heath. Here are the remains of an ancient abbey of Bernardines, founded in 1170, for which Ada Counteſs of March ſwore fealty to Edward I. of England, in 1296. The Earl of Wemyſs has lately built an elegant ſporting villa, called the *Retreat*, about a mile from the ſmall Kirktown of *Abbay*. In 1801, the population was 138; decreaſe 61 ſince 1791, but an increaſe of 48 ſince 1755.

ABBEY PARISH of PAISLEY. Vide Paisley.

ABBEY-GREEN; a village in Lanarkſhire, in the pariſh of Leſmahagoe; ſo named from being built around the remains of a monaſtery, dependent on that of Kelſo, founded by King David I. in 1140, and dedicated to St. Macule. All that remains of the monaſtery is a ſquare tower, with battlements, now converted into a ſteeple to the church of Leſmahagoe. The village lies 12 miles S. of Hamilton, and 4 from Lanark. In 1793, it contained about 95 houſes, and 430 inhabitants.

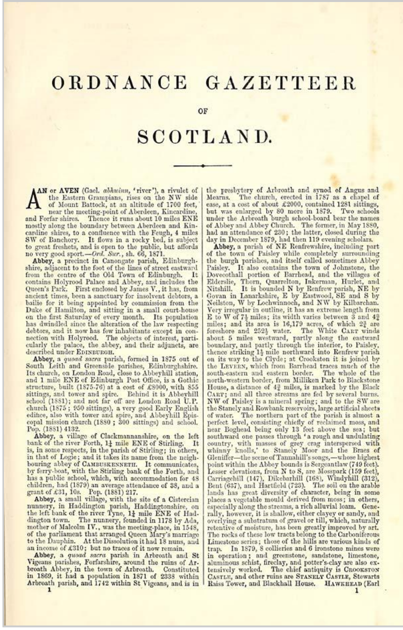
ABBOTRULE; a pariſh in Roxburghſhire, lately ſuppreſſed, and divided between the pariſhes of Bedrule and Southdean.

Abb's (St.) Head; a promontory of land, well known by ſeamen, lying in the pariſh of Coldingham, county of Berwick, about 10 miles N. of Berwick, and the ſame diſtance S. of Dunbar. Longitude 1° 56′ W. latitude 55° 54′ N.

ABBOTS-HALL; a village and pariſh in the S. coaſt of the county of Fife. It is a ſmall irregular pariſh, being in its utmoſt extent not more than 2 miles each way. Its general appearance is very pleaſant, riſing gradually from the coaſt northward, into pretty high grounds. The ſoil is thin, but exceedingly fertile, particularly in warm, ſhowery ſummers. Mr. Ferguſon of *Raith*, one of the chief proprietors of the pariſh, has lately made out ſome extenſive plantations round his ſeat, and erected a fine obſervatory on the higheſt ground in the pariſh, which has a very commanding proſpect. The diſtrict abounds with coal and limeſtone; in

A

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 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.

All resources presented in tis work 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 corpus computationally interoperable, queryable, and reusable across domains. / The remainder of this paper is structured as follows. Section 2 reviews related work on Semantic Web methods for cultural heritage, including digitization, ontology design, and the use of large language model. Section 3 provides a detailed description of the MappingChange resource. Section 4 details the Heritage Textual Ontology. 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 (ontolgy available here), 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.

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 and METS XML for thousands of 19th-century documents, including the Gazetteers of Scotland. However, 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 the Heritage Textual Ontology (HTO [7]), 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 [8] 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, this work is fully integrated into the Frances [9] semantic platform, which supports temporal exploration and semantic search of historical data. Recent updates to *Frances* include improved support for collection browsing, semantic and full-text search, concept evolution over time, and geolocation visualization, 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. Each DataFrame is normalized to a one-entry-per-row format, making it suitable for downstream integration and statistical analysis. We provide both individual DataFrames for each of the ten editions and a unified, cross-edition DataFrame that aggregates all entries into a single structure.

These DataFrames serve as the foundation for constructing 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 includes pre-computed sentence embeddings for each article and 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 curated set of Jupyter notebooks. These include SPARQL query examples and data exploration analyses. The three interlinked knowledge graphs are deployed within the Frances semantic platform, providing users with interactive interfaces for querying, timeline visualization, and concept exploration.

The full resource (including individual and aggregated DataFrames, knowledge graphs, and exploratory notebooks) is openly available at MappingChange repository, with persistent identifiers to be issued via Zenodo. To promote reuse and reproducibility, the repository includes modular scripts and comprehensive documentation. The step-by-step process by which the data and knowledge graphs are constructed—including article extraction, enrichment, and RDF serialization—is described in Section 5.

4. Heritage Textual Ontology (HTO)

The HTO provides the semantic backbone for *MappingChange*, enabling the structured representation of historical textual records, their provenance, and the evolving concepts they describe. Developed to support real-world use cases in digital heritage, HTO models not only entities and attributes, but also the editorial and computational processes by which historical texts are extracted, digitized, interpreted, and semantically enriched. Unlike more generic vocabularies, HTO is tailored to the challenges of heritage corpora—such as OCR noise, editorial variation, and evolving terminology—offering fine-grained support for provenance, textual quality, named entity recognition, and diachronic conceptual alignment.

HTO is openly developed, and it builds on established ontologies such as PROV-O, SKOS, Schema.org, and CIDOC CRM, while introducing domain-specific classes and properties designed for flexible reuse and extension. Since its initial release, HTO has been extended to support geospatial annotation, richer NER-based enrichment, and explicit modeling of language-model-based transformations, making it suitable for a wider range of digitized corpora.

HTO is modular and extensible, and can be adopted in other projects that require modeling of OCR-derived documents, provenance-aware digitization pipelines, or diachronic semantic alignment. It also plays a central role in enabling FAIR knowledge graph construction with transparent lineage tracking.

4.1 Modeling Bibliographic Structure and Provenance

HTO provides a structured vocabulary for modeling the archival hierarchy of heritage texts (including works, editions, series, and pages) through the class `hto:Work` and its subclasses. Each textual entity is also a `prov:Entity` (`hto:EntityWithProvenance`), enabling provenance tracking for digitization methods (e.g., OCR or manual transcription), quality levels (`hto:TextQuality`), and attribution to specific software or human agents (`prov:Agent`). Works can be grouped into collections using `hto:Collection` and linked to their physical or digital source editions. Figure 2 illustrates this bibliographic modeling layer.

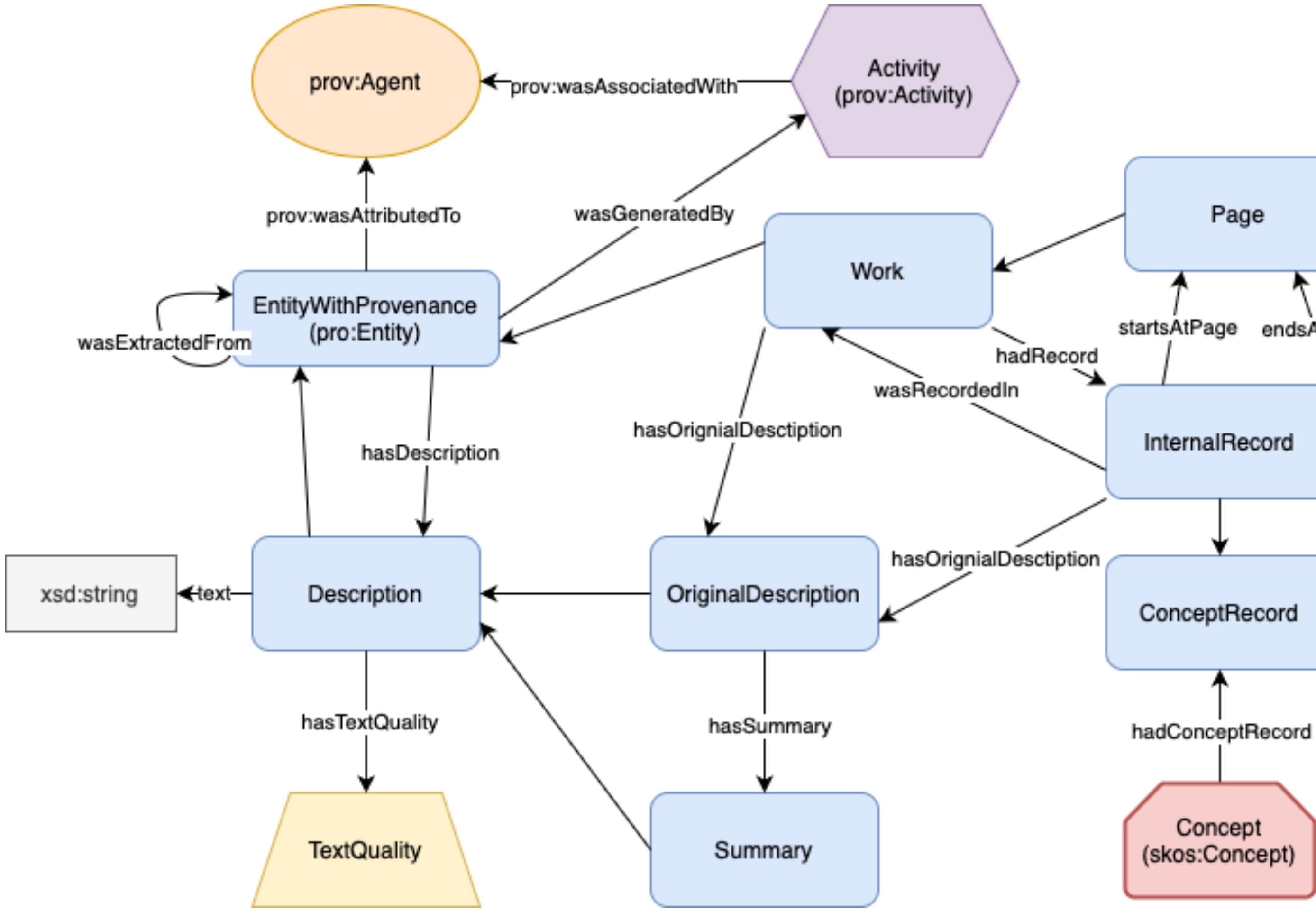


Figure 2: Bibliographic and provenance modeling in HTO, including core classes (blue), agents (orange), locations (green), and datatype properties (grey).

4.1 Capturing Textual Records and Interpretations

HTO distinguishes between original descriptions (`hto:OriginalDescription`) and derived or enriched interpretations (`hto:Description`). These are associated with their source pages via `hto:Page` and `hto:hasOriginalDescription`, and annotated with quality levels (e.g., “Low”, “Moderate”, or “High”) based on their provenance. This allows the knowledge base to retain multiple text variants and trace how they were extracted and transformed.

Records within the text are modeled using `hto:InternalRecord` (for local entities) and `hto:ExternalRecord` (for linked data resources such as Wikidata or DBpedia). Repeated terms (like “St Andrews”) are tracked across editions via `hto:TermRecord`, while `hto:ConceptRecord` groups semantically similar entries into shared `hto:Concept`s, enabling diachronic alignment. Figure 3 shows how textual terms and their semantic clusters are represented.

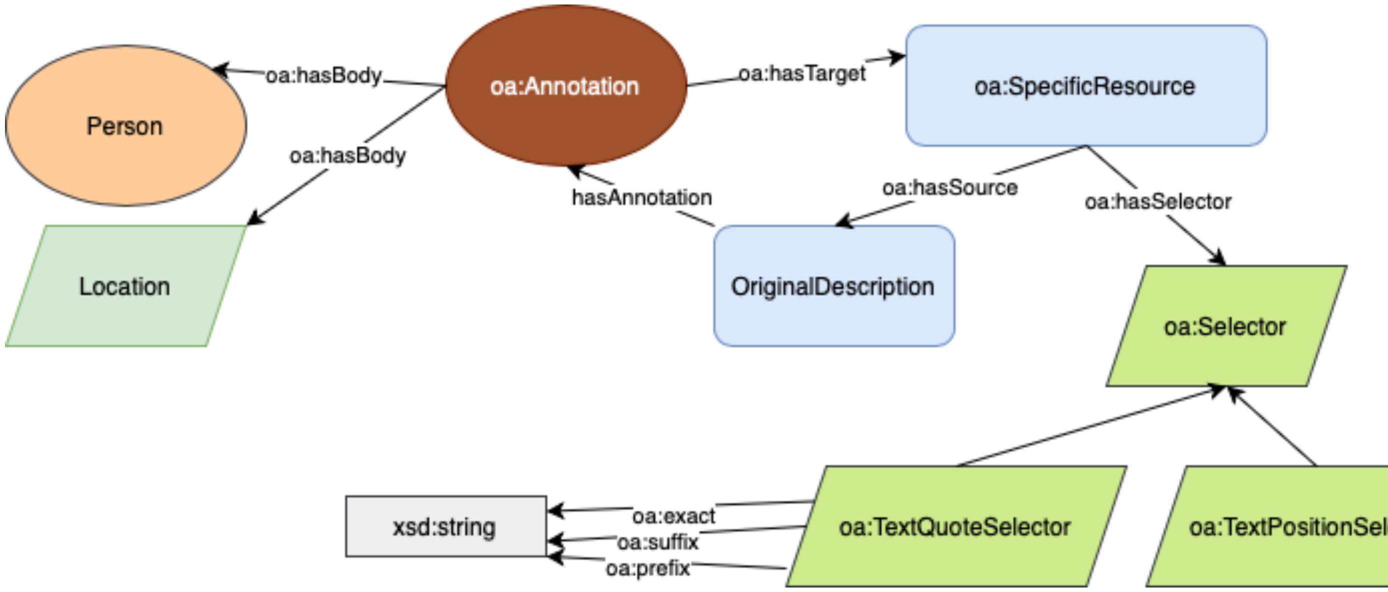


Figure 3: Modeling of term records and concept clusters using HTO classes for internal and external alignment.

4.1 Geospatial Annotations and Place Modeling

In *MappingChange*, HTO has been extended to support spatial annotation and georesolution (See Figure 4). Locations are represented as `hto:Location` and typed using subclasses such as `hto:Town`, `hto:Region`, or `hto:River`. Each place may include spatial geometries using `geo:hasGeometry` or `geo:asGeoJSON`, with optional declarations using `cidoc-crm:SP2_Declarative_Place`. This enables spatial reasoning, integration with GeoSPARQL, and alignment with modern gazetteers and linked data services. or an overview of spatial modeling in HTO.

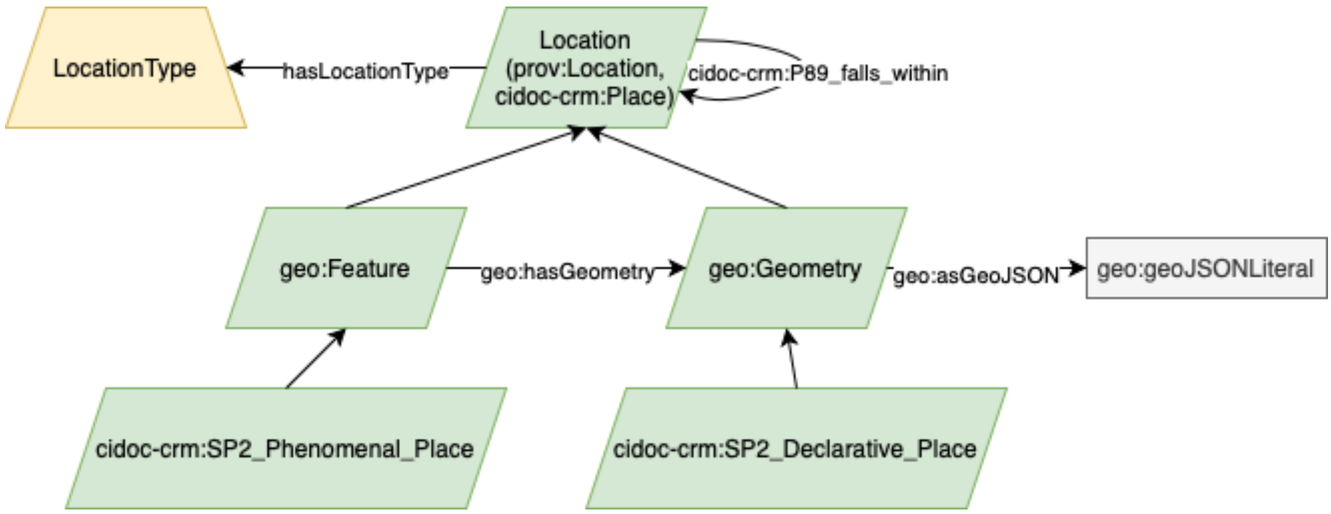


Figure 4: Representation of locations and spatial types in HTO, including georesolved coordinates and place categorization.

4.1 Annotation and Segment Selection

To link extracted place names or descriptions to their textual anchors, HTO adopts Web Annotation (OA) standards. Each annotation (`oa:Annotation`) includes a source document, a target entity, and selectors (e.g., `oa:TextQuoteSelector`, `oa:TextPositionSelector`) that specify the exact span of text involved. This ensures fine-grained traceability back to OCR-aligned sources and supports downstream validation and curation. An example is shown in Figure 5.

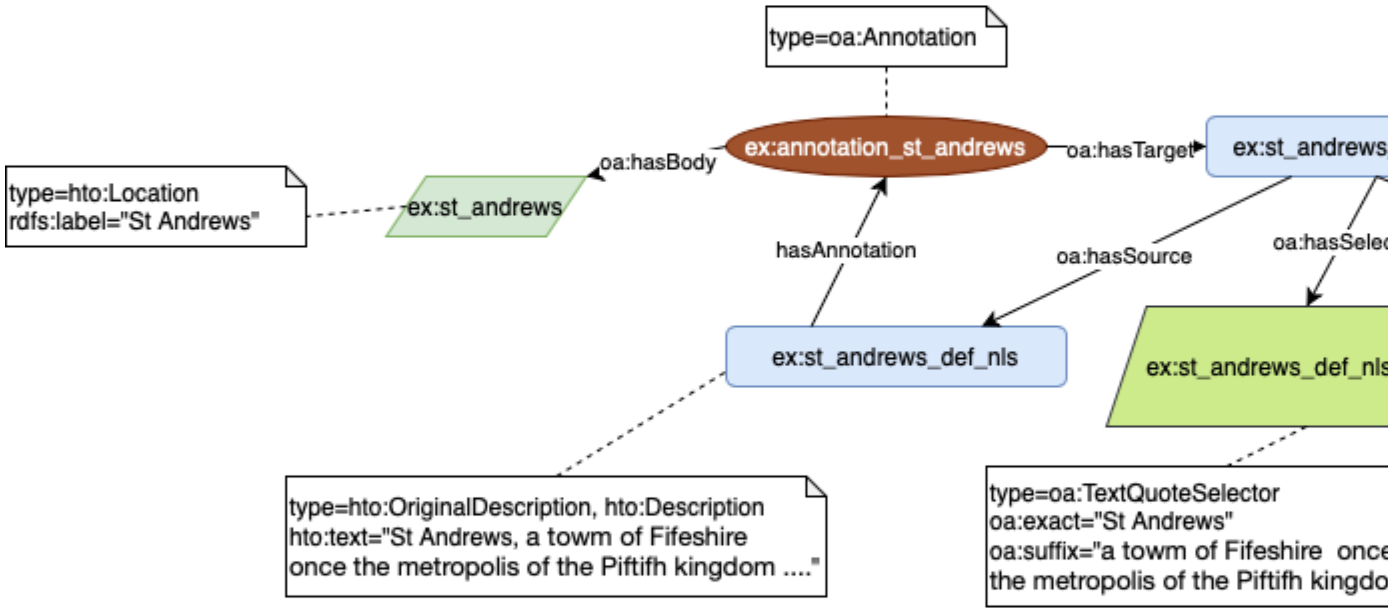


Figure 5: Example of how text segments are annotated and anchored using OA selectors in HTO.

5. Construction and Content

The MappingChange knowledge base is built through a multi-stage pipeline that transforms unstructured OCR-aligned ALTO XML files into structured DataFrames and RDF graphs. Each stage of the pipeline is modular and reproducible, with edition-specific scripts documented and openly available in the MappingChange GitHub repository.

Of the twelve editions in the NLS collection, we process ten (See Figure 6) as fully descriptive gazetteers with complete metadata and 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 is available. These ten editions form the basis of the MappingChange resource, with each transformed into a structured DataFrame containing one row per article-level entry.

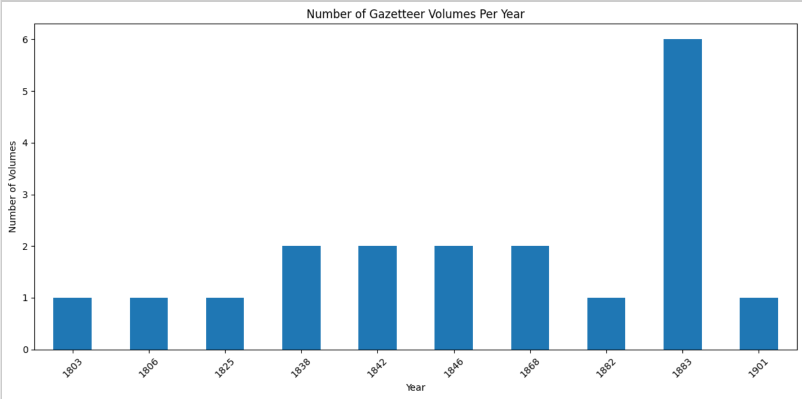


Figure 6: Number of volumes per gazetteer edition (1803–1901). The 1883 edition spans six volumes, while most others are single- or double-volume works.

5.1 Article Extraction and Prompt Engineering

As input to our pipeline, we use the gazetteers_dataframe, a cleaned and consolidated DataFrame derived from our earlier Gazetteer_HTO knowledge graph [10]. This resource contains entries from the gazetteers editions, each representing the full OCR text of a single page, along with metadata such as edition identifier, volume, page number, and candidate place names.

The goal of this stage is to segment each page into distinct articles and extract structured place descriptions. This task presents several challenges: (a) place names can be ambiguous and refer to multiple locations across different editions; (b) many descriptions span multiple pages; (c) new places are introduced in later editions; (d) descriptions frequently reference other places (e.g *ABBEY PARISH*. *See Paisley*.), which must also be identified; and (e) places may be listed under multiple or alternative names. Our extraction scripts and prompting strategies are designed to address these complexities by isolating each named place and its corresponding description, including cross-references and aliases.

At the core of the article segmentation process is a set of Python scripts (`extract_gaz_*.py`) that apply GPT-4 to extract article-level place descriptions from OCR text. Each edition has its own script to accommodate layout, editorial, and typographic idiosyncrasies—including abbreviation handling, redirects, and multi-page articles.

A key part of this process is the design of GPT-4 prompts tailored to each gazetteer's editorial conventions. The base prompt follows an instruction format asking the model to identify and extract gazetteer articles from a page of OCR text. Variants were introduced to handle structural nuances—such as distinguishing index lines from articles in early editions or handling mid-page redirects in later ones.

The table below summarizes the edition-specific characteristics and adaptations in the prompting strategy:

Edition	Layout/Format Features	Prompt Adjustments
1803	All caps titles, minimal punctuation, mid-column entries	Prompt includes rules for semicolon-delimited entries
1806	Similar to 1803 with improved spacing	Added regex pre-filters to exclude 3-letter headers
1825	Shorter entries, denser formatting	Emphasis on short entries and abbreviation disambiguation
1838	Two-column format, clearer title separation	Prompt refined to distinguish article breaks explicitly
1842	Redirects common, layout noisy	Includes logic for disambiguating abbreviated redirects
1846	Continued abbreviation patterns, multi-page entries	Includes continuity checks and redirect expansion
1868	Longer, structured entries with location hierarchies	Added cues for nested article types and locations
1884–1901	Title-cased entries, structured and clean layout	Simplified prompts; uses typographic features directly

All scripts tokenize OCR text by page and apply the prompt in a batch-efficient manner. Outputs are parsed into structured JSON records that include article text, place name, page number, and other metadata. These are aggregated into edition-specific DataFrames.

5.1 DataFrames to RDF

Each JSON DataFrame is cleaned and mapped to RDF using the Heritage Textual Ontology (HTO). For each article, we instantiate one or more `hto:Description` entities, annotated with quality metrics, extraction method (GPT-4), and source metadata (edition, volume, page). This conversion step is implemented using reusable mapping scripts in Python and SPARQL.

5.1 Semantic Enrichment and Linking

Following RDF generation, the dataset undergoes semantic enrichment through three main procedures:

- **Concept Clustering**: Sentence embeddings and clustering algorithms group semantically similar articles across editions. Resulting `hto:Concept` instances link equivalent or evolving descriptions of the same place.

- **Entity Linking**: `hto:PlaceRecord` and `hto:Concept` instances are matched to external resources via Wikidata and DBpedia using a hybrid of embedding-based and string-matching techniques.
- **Geospatial Annotation**: Named Entity Recognition (NER) and georesolution tools (e.g., Edinburgh Geoparser) are used to annotate and disambiguate place mentions. Each resolved location is stored as a `hto:GeographicAnnotation` with coordinates, spatial type, and provenance.

5.1 Knowledge Graph Serialization and Deployment

The final RDF outputs are serialized in Turtle and hosted in a public Fuseki SPARQL endpoint. In parallel, Elasticsearch indices are built from the JSON and RDF data to support keyword search and semantic similarity queries using embeddings.

All steps—from GPT-4 prompting to RDF generation—are documented in executable scripts and notebooks in the GitHub repository, ensuring full reproducibility. Each transformation stage is also recorded in HTO using provenance properties (`prov:wasGeneratedBy`, `hto:hasTextQuality`, etc.), enabling end-to-end traceability of every knowledge graph triple.

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

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