

# 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

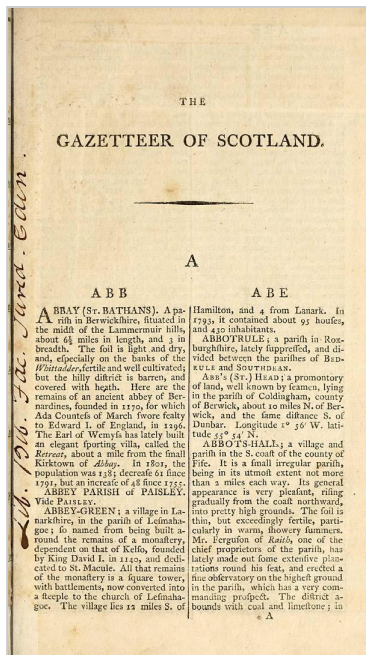
Descriptive gazetteers were a cornerstone of how 19th-century Scotland recorded and transmitted its geographical knowledge—capturing towns, parishes, rivers, castles, and natural features within evolving historical, social, and economic narratives. As industrialization, migration, land reform, and empire 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”) 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 Elasticsearch indices) 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.



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.

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 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, this 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 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 (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.

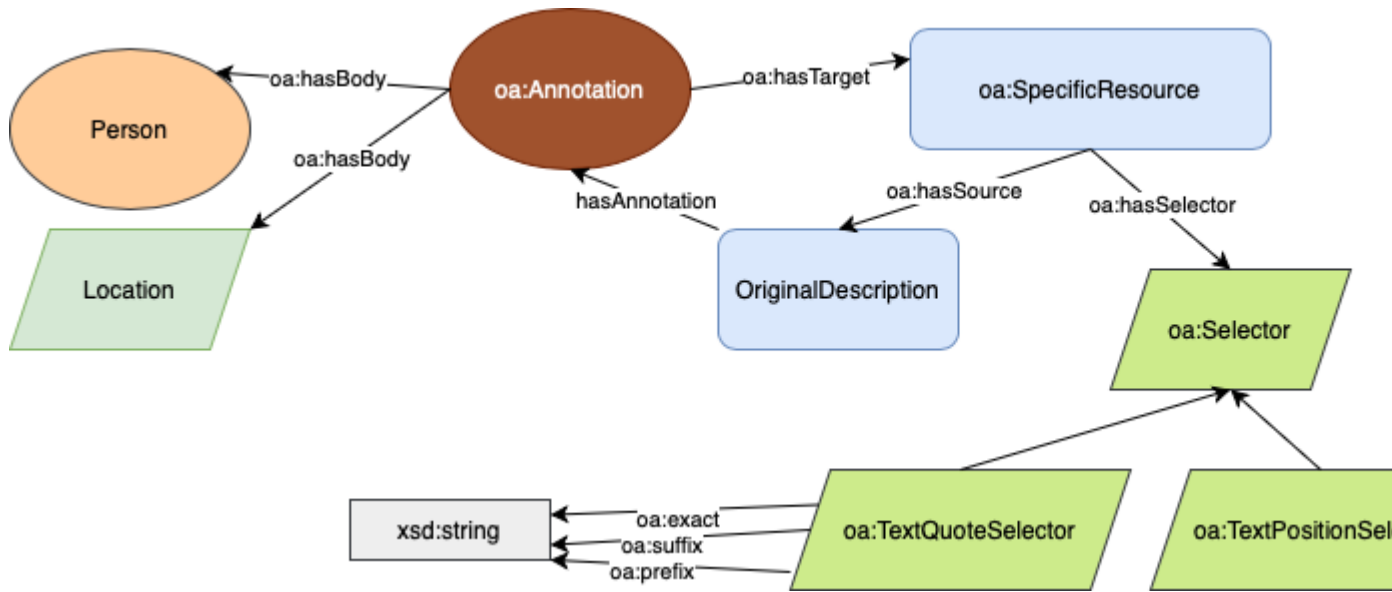


**Figure 2:** Bibliographic and provenance modeling in HTO, including core classes (blue), agents (orange), locations (green), and datatype properties (grey).

#### 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.

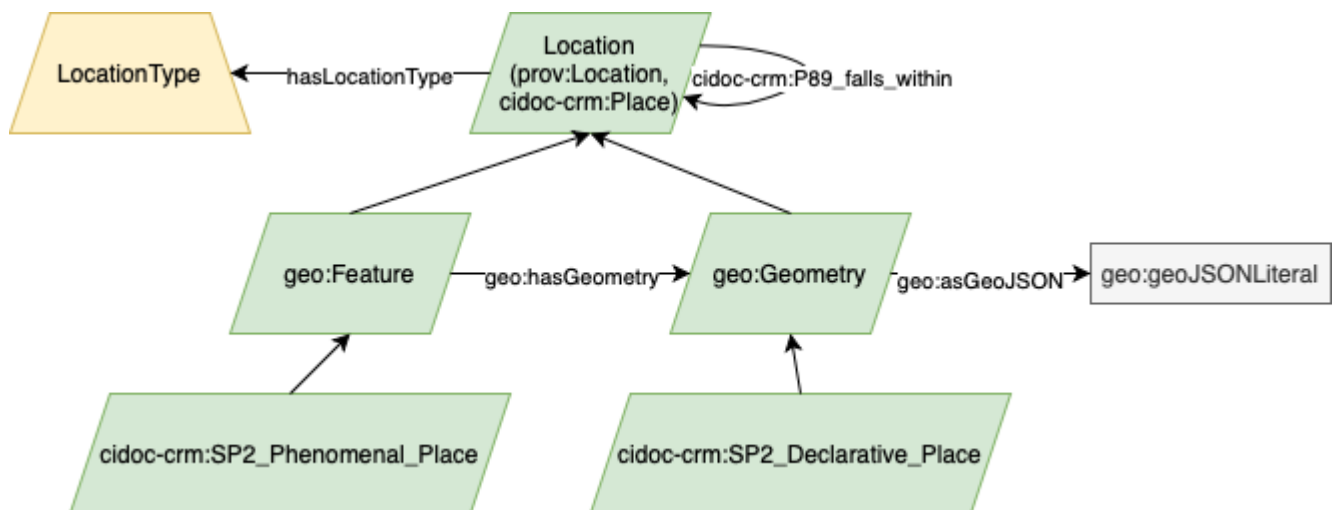
Records within the text are modeled using `hto:InternalRecord` (for local entities) and `hto:ExternalRecord` (for linked data resources such as Wikidata or DBpedia). Repeated terms (like “St Andrews”) are tracked across editions via `hto:TermRecord`, while `hto:ConceptRecord` groups semantically similar entries into shared `hto:Concepts`, 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.

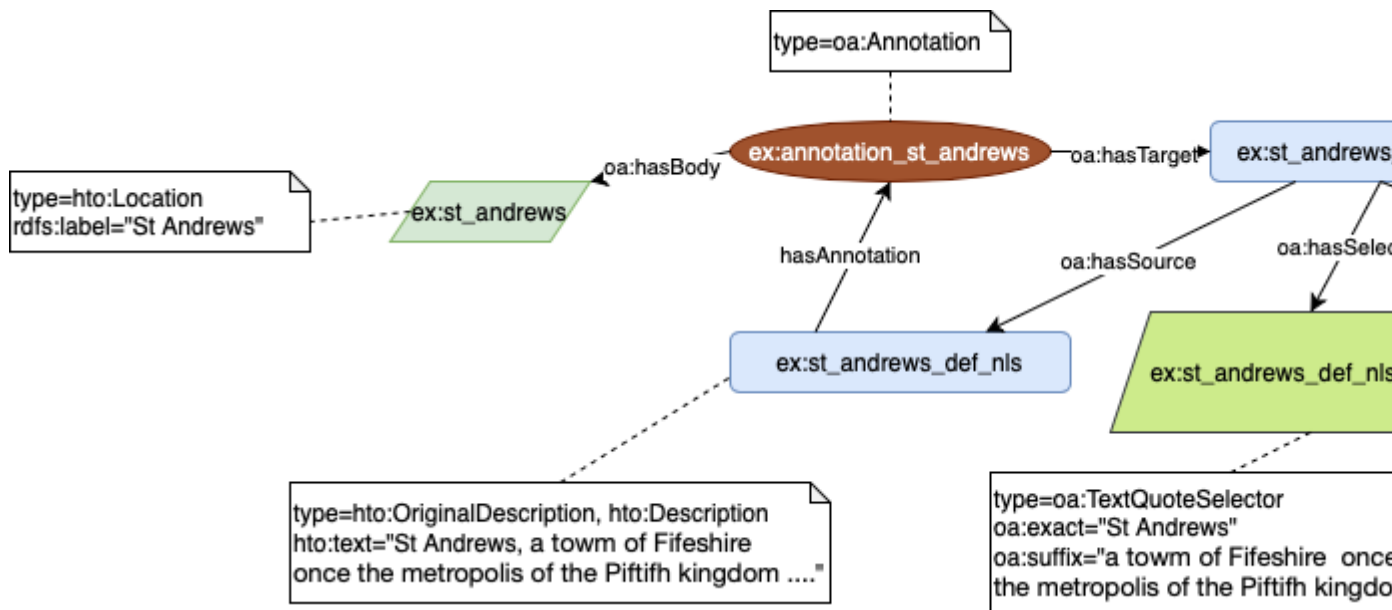


**Figure 4:** Representation of locations and spatial types in HTO, including georesolved coordinates and place categorization.



## 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 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:

<b>Gazetteer Volume</b>	<b>Prompt Focus</b>	<b>Format Characteristics</b>	<b>Prompt Adaptation Strategy</b>
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

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