A Cypher Knowledge Graph of British Imperial Territorial Evolution

Wikidata and teamwork without teams in computational history

Jim Clifford

2026-05-14

This paper presents a curated Cypher knowledge graph documenting the territorial evolution of the British Empire from the early-modern period through the late twentieth century, alongside an interactive Sankey-style visualization of colonial successions, partitions, federations, and independence transitions. The graph contains 747 historical territories (314 colonial polities and 433 princely states) connected by 1,203 typed relationships, with every entity grounded to a Wikidata QID. The argument is twofold. First, property graphs in the Cypher query language offer a productive substrate for political-territorial history: they make the messy genealogies of empire (partitions, reorganizations, post-imperial successor states) explicit and queryable without forcing them into the cleaner abstractions of relational schemas, while leaving room for stricter ontologies like CIDOC-CRM where the work demands them. Second, grounding historical entities to Wikidata QIDs creates shared foundations across projects, allowing computational historians to do teamwork without forming teams. The paper also documents the iterative, human-in-the-loop verification workflow used to keep hallucinated QIDs out of the graph.

1 Introduction

Historiography is cumulative: each new argument depends on prior scholarship and primary sources that have named, dated, and located the same actors and places. The structured datasets historians now produce should build upon each other in the same way. With coding agents making it much easier to create historical datasets, we need to reinforce linked open data conventions among computational historians to make our data interoperable. The prerequisite is shared ground: agreement that two records describing "Victoria" or "Huizong" or "Demerara" describe the same entity. For prose scholarship, that ground is supplied by citation. For structured data, it has to be supplied by persistent identifiers. At the scale historical research now operates, Wikidata, with over 120 million items including millions of historical figures and places, is the only identifier system with sufficient coverage to play that role. This paper presents a Cypher property-graph dataset of 747 historical territories and 1,203 typed relationships, together with an interactive Sankey visualization that makes the genealogies legible.

This dataset is a foundation for research on the history of British imperialism and the British World System between the sixteenth and twentieth centuries. The goal was a database recording all of the colonies in any given year along with the evolution of individual colonies throughout their history. Most of this information already existed in Wikipedia and its linked open data project Wikidata. Working with Claude Code, I grounded the territories to Wikidata QIDs for persistent identifiers. Early in the project, however, it became clear that coding agents regularly hallucinated Wikidata QIDs, so the work required extensive human-in-the-loop verification. I found that developing a visualization helped with that verification by making it easy to see whether the colonies were linked together correctly. This paper provides an overview of the iterative process used to develop and check the graph with Claude Code.

Cypher is the query language used by graph databases such as Neo4j and FalkorDB; I chose a property-graph model over the more standards-bound RDF/SPARQL stack because relationships in a property graph can carry their own properties (a date, a source, a qualification), which matters when the relationship itself (a partition, a federation, an independence transition) is the historical object of interest, not just a link between two things. The dataset is a basic starting point, ready to be extended with additional data ranging from trade statistics to nodes for political officials or key events. The argument is larger: that knowledge graphs, with entities grounded to Wikidata QIDs, offer a productive *shared foundation* for historical research at this scale. Anyone who takes this data, builds on top of it, and shares the results contributes to a larger historical project. It allows us to do teamwork without forming teams.

2 Wikidata grounding as a shared foundation

Computational history is now possible for individual scholars without grant-funded labs. Coding agents are empowering thousands of historians to build new datasets, but this risks a proliferation of data silos. If we all assign different unique identifiers to disambiguate Victoria the Queen from Victoria, British Columbia, we end up rebuilding the same disambiguation work in every project, and losing the chance for our datasets to compose into anything larger than themselves.

This is the problem linked open data was designed to solve. The idea is straightforward: if every dataset attaches the same persistent identifier to the same entity, then datasets compose. Wikidata has already done much of this work. The first page of results for "Victoria" disambiguates the queen, the British Columbian capital, the Australian state, the capital of the Seychelles, a Roman goddess of victory, a main-belt asteroid, the Victoria and Albert Museum, and roughly a dozen others, each with a permanent identifier any historian's dataset can attach to.

Consider what this means in practice. A historian working on nineteenth-century Nova Scotia and a historian working on the East India Company army are unlikely to know each other's work. Their corpora are different, their archives are in different buildings, and their conferences do not overlap. But if both of them ground their data to Q335381, their datasets join automatically, even though one historian's records name him George Ramsay, Governor of Nova Scotia from 1816, and the

other names him the Earl of Dalhousie, Commander-in-Chief in India from 1830. A query about Dalhousie's career can now reach across one historian's biographical work in Halifax archives and another's regimental records in Calcutta, without either of them having planned for the integration. The work that previously required weeks of reconciliation becomes a single query. The promise of computational history is not that everyone produces their own dataset. It is that many small datasets compose into a research commons larger than any of them.

There have been concerns in the digital humanities community about using Wikidata as scholarly infrastructure. Wikidata is crowdsourced; its data model is determined by its community of editors; some of its modelling decisions, of which the handling of gender is the canonical case, have been incompatible with the priorities of researchers working on people whose lives the standard categories misrepresent. The response, in projects like LINCS (Linked Infrastructure for Networked Cultural Scholarship), was to build proper scholarly infrastructure with controlled vocabularies, considered ontological commitments, and editorial accountability. Those choices were correct, and I remain hesitant about building knowledge graphs on the Wikidata platform itself.

But the worry applies to *hosting* scholarship on Wikidata, treating Wikidata itself as the place where your scholarly conclusions live, where your data sits, and where your interpretations are subject to revision by anyone who edits the relevant pages. It does not apply to *grounding* scholarship to Wikidata, which is a different operation. LINCS already works this way. The graph LINCS publishes is its own linked open data, carefully modelled, with editorial accountability to the scholars who contribute to it. LINCS uses Wikidata QIDs directly as the identifiers for entities Wikidata already covers, and mints its own only when no alternative exists. The model is sovereign; the identifiers are shared. Computational historians can choose between fully compliant CIDOC-CRM linked open data and simple research spreadsheets. If we attach Wikidata identifiers to the entities in our data, we make our work interoperable with other projects. Anyone working on a related corpus, yours or someone else's, now or in twenty years, can join the two without asking permission, because the identifier is shared.

The reason this matters is that the alternative does not scale. No scholarly project, no matter how well-funded, can produce identifiers for over 120 million historical people, places, organisms, events, and concepts on its own. Wikidata has done that work. Its coverage is uneven and its modelling has problems, but it is the only system at the right scale for historical research as actually practised. Ignoring it produces silos. Grounding to it, while keeping your own dataset, allows historians to take advantage of the strengths of Wikidata without incorporating its problems. A historian working on Bengal in 1850 and a historian working on Jamaica in 1840 do not need to coordinate, share infrastructure, or even know each other to produce datasets that interoperate. They need to use Wikidata identifiers for the people, places, and institutions in their sources. That is teamwork without a team. This is not a call for centralized infrastructure. It is the opposite. It is a call for a small set of shared conventions. This lets individuals and small groups produce work that connects to a research commons without anyone having to build the commons explicitly. The commons is the consequence of the convention.

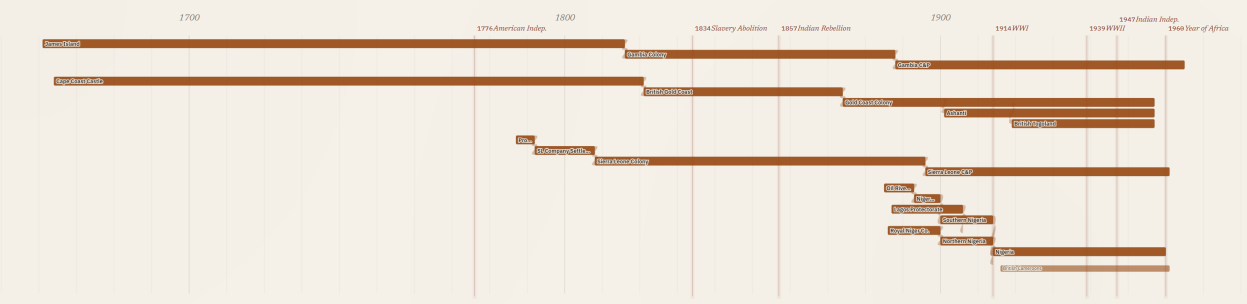
As the remainder of this working paper will demonstrate, the process is relatively easy and requires three basic commitments:

- First, add Wikidata identifiers to the entities in your datasets. If you have a spreadsheet of historical figures, add a column for their Wikidata identifier.
- Second, publish your data with the identifiers attached. A project website, a Zenodo deposit, a GitHub repository: the venue matters less than the principle that the identifiers are present and durable, so that someone reading your work in ten years can still find them.
- Third, treat Wikidata-editing as part of historical practice. When you find that a colonial administrator, a rural township, or an eighteenth-century concept lacks an entry, contribute one with sources. We need to make this into an evolving standard of methodological rigour, on the same continuum as proper archival citation.

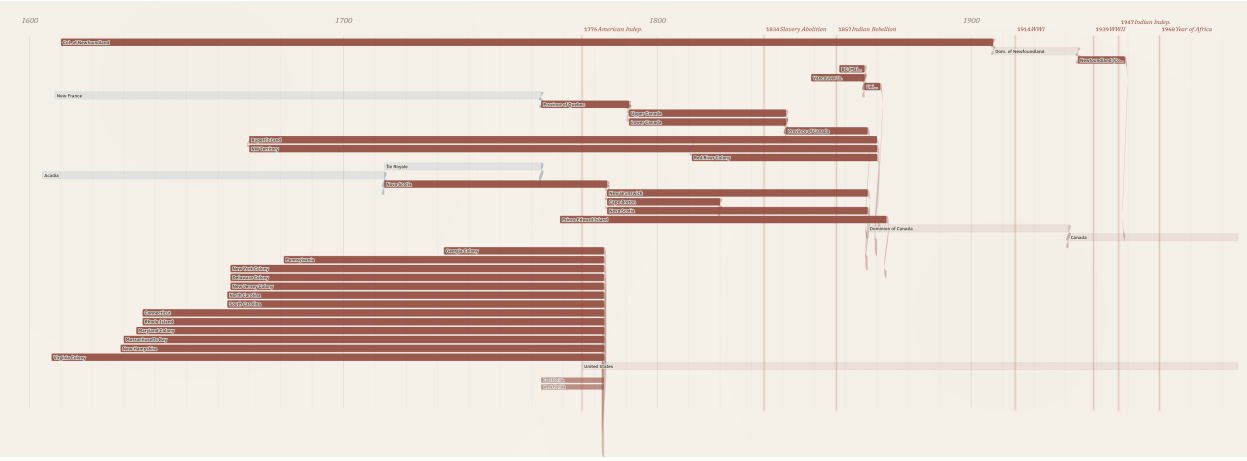
2.1 The British Empire Evolution Knowledge Graph

This project started small and spiraled out of control. I was testing LLMs' ability to extract and ground information from Internet Archive documents related to the British Empire in the nineteenth century. It occurred to me I should start with a foundational dataset of all the colonies. I asked a coding agent to help me create a knowledge graph of all the colonies in the late nineteenth century. It quickly produced a graph that covered a lot of the empire, but was riddled with errors. So the project snowballed as I tried to figure out how to improve accuracy and extend the scope to include the full overseas empire. Wikidata was an obvious starting place. I knew from the Trading Consequences project (2012-2016) that GeoNames does not include many of the historical geographical entities such as Rupert's Land or the Oil Rivers Protectorate. Wikidata, in contrast, has all of these entities as it is linked to Wikipedia articles. Most British colonies have a Wikipedia entry and a Wikidata ID: https://en.wikipedia.org/wiki/Niger_Coast_Protectorate and https://www.wikidata.org/wiki/Q2566427. Wikidata also has entries for minor princely states without an English language Wikipedia page, such as Dedhrota (https://www.wikidata.org/wiki/Q131126101). Working with Claude Code, I tried to identify and ground all of the British colonies in Wikidata and develop them into a knowledge graph using Neo4j as the graph engine. The project grew to include colonies starting with the Crown of Ireland Act of 1541 and the founding of Virginia in 1607. The project was iterative as I identified errors and worked with Claude to develop plans to verify the data. I realized I needed to be able to see the colonies and the temporal relationships between colonies as they merged, split or gained independence. This pushed me to develop a visualization to make the data easier to follow and correct.

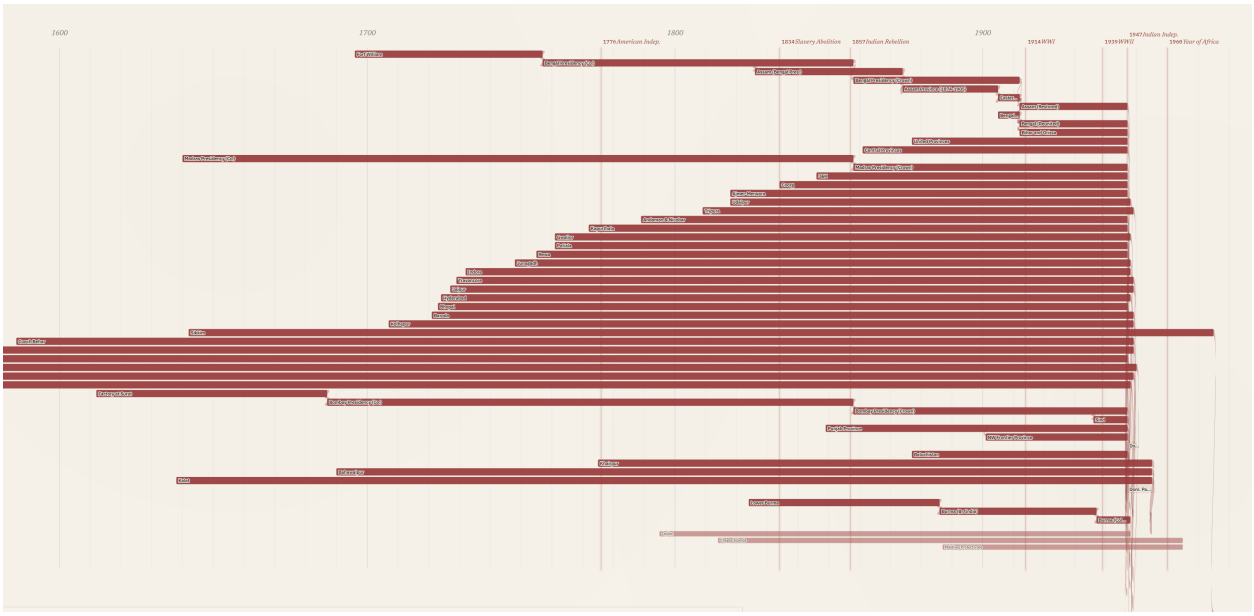
The visualization needed to track change over time and show the chain of relationships as smaller coastal colonies merged and evolved into larger colonies, as was the case in West Africa.



The visualization proved important in correctly mapping out the colonial history of what became Canada, capturing the conquest of the French Empire, the division and later merger of Upper and Lower Canada, and the shorter history of two colonies merging into one before British Columbia joined Confederation.



South Asia created unique challenges: hundreds of princely states, including many that predated Britain's overseas empire. I decided to select a geographically diverse sample of these states in the visualization while including all of them in the underlying knowledge graph. The South Asia section highlights a limit in this visualization, where screen space and data constraints meant every colony is represented by the same width of bar, so the Bombay Presidency appears the same as the Factory at Surat despite their different scale and levels of colonial control. I could use population estimates or square kilometres of territory to differentiate between the major colonies and minor trade forts, but that would require data for each colony and a way to represent change over time. The current visualization prioritizes identifying all of the colonies, their time span and how they relate over visualizing the social, political or economic significance of the different colonies.



The iterative process continued after completing the visualization. I needed to ensure all of the Wikidata QIDs matched the colonies. Most of them were reviewed in an earlier human-in-the-loop process. I used Claude Code to create an HTML website that pulled the description from Wikidata and presented it alongside the colony name. I then either verified or provided an alternative. As I worked through the list, I found some minor island colonies did not have a Wikidata ID for the colonial period, so I decided to use the ID for the geographical islands. I also found a few colonies not included in Wikidata and had to decide whether the current political entity could be used as a stand in. In a final process, I used Claude to search every row in the knowledge graph to ensure the Wikidata ID matched, and it found a small number of errors. I completed a human review of these errors and published the final version with notes attached for all of the colonies where I had to use a geographical entity, the modern political entity, or a broader Wikidata entity that spans multiple sub-periods of colonial administration, in place of a unique Wikidata ID for the colony. I am now considering adding these entities into Wikidata as a contribution to that project and to make the knowledge graph more consistent.

2.2 Property graphs in historical research

The choice to model in Cypher places this dataset within a small but growing tradition of historical and digital-humanities projects that have made the same call. Christopher Warren and colleagues built *Six Degrees of Francis Bacon* (Warren et al. 2016) on a property graph of roughly 13,000 early-modern persons and 200,000 typed relationships (PARTNER_OF, LETTER_TO, TRANSLATED, LIT_CRIT_ON), and documented an evolution from storing relationships as properties of nodes to adding dedicated nodes and labels as their research questions matured. The same model has been applied across very different historical subfields, including Roman prosopography (Varga and Bornhofen 2024), network analyses of postwar reparations activism (Pan 2022), and medieval Korean kinship and patronage networks (Cha 2026). Jon MacKay's (2018) *Programming Historian*

tutorial, built around the 1912 Canadian corporate-interlock directorates, has made Cypher part of the field's pedagogical apparatus.

What unites these projects is not a vendor preference but a recurring set of fits between the property-graph model and the work historians actually do. Edges that carry their own properties let scholars annotate ties with dates, sources, and qualifications without distorting the schema around them. Questions that follow a chain of relationships across many steps, such as a political succession from colony to independent state to federation, or a kinship descent across generations, can be asked of the database in a single query rather than reconstructed step by step from many smaller ones. And because the schema can evolve as the source material is better understood, the model does not require committing to an ontology before the research question is fully formed, which fits the iterative way historical understanding develops.

Where the field is still thin is in published humanist work on GraphRAG, the LLM-augmented retrieval pattern that uses a knowledge graph as both context for and constraint on generation. Computer-science work on GraphRAG has accelerated through 2024 and 2025 (Edge et al. 2024; Han et al. 2025), and cultural-heritage-adjacent applications have begun to appear, but a flagship humanist-authored study of GraphRAG on historical archives has not yet been published. This is an opening rather than a settled question. Property-graph datasets like this one are well-positioned for that work when historians take it up, and the present paper is intended in part as a substrate that GraphRAG-oriented research can attach to.

2.3 Property graphs now, RDF later

A knowledge graph can be structured in more than one way, and the choice matters enough that a historian should understand it before adopting either. This project is a property graph: territories are nodes, the relationships between them are labelled edges, and both carry properties such as dates and sources. The main alternative is to publish the same material as linked open data, using a storage format called RDF and an international standard vocabulary, CIDOC-CRM, designed for cultural-heritage data. CIDOC-CRM is carefully designed and the most widely adopted shared vocabulary of its kind in the digital humanities, and it is built for exactly the cross-project composition this paper argues for: material modelled in it can in principle be combined with any other project that has done the same. It is reasonable to ask why I did not use that standard from the start. The answer is not that it is the wrong target. It is that a property graph is far easier to build and to read, and the standards-based version can still be produced from it at the end.

It helps to be concrete about what a CIDOC-CRM version of this dataset would involve. Each polity would become a persistent entity node. Each of the transitions this project records as a typed edge (a colony evolving into a dominion, a territory partitioned into successor states, several colonies federating into one) would become an event node, carrying its own date, its own provenance, and its own type.[1] The relationship a historian actually wants to query, "A became B," is then never

[1] In CIDOC-CRM terms, each polity is a persistent entity such as an E74 Group; the transitions are E81 Transformation events, with E66 Formation and E68 Dissolution for entry into and exit from existence, each with an

an edge to follow but a path through an intervening event. The 1,203 typed relationships in this graph would become 1,203 event nodes and their scaffolding, before anything new is said about the empire.

A property graph is easier to build because the model maps directly onto what the historian is asserting. "The Province of Canada became the Dominion of Canada in 1867" is one edge with one date. There is no prior decision about which event class applies, no time-span node to instantiate, no property whose domain and range have to be checked. This matters most for the way this project was built: a single historian working with coding agents. I can ask Claude Code to produce a CIDOC-CRM model, and it will produce one. What I cannot do is be confident it is correct. CIDOC-CRM is a large and tightly specified ontology; a leading LLM struggles to apply all of its rules consistently, and a modelling expert can reliably say what a draft gets wrong but not as reliably certify what is right. The property graph is not just the thing this workflow can build. It is the thing it can build and trust.

A property graph is also easier to read. The graph as drawn is the genealogy: a reader sees the Province of Quebec become the Province of Canada become the Dominion of Canada, each step a labelled edge. In the CIDOC-CRM version the genealogy is still there, but it is implicit in chains of event nodes that the reader has to know how to traverse. For a dataset whose whole contribution is a queryable, legible genealogy, that directness is not a convenience. It is the point.

The honest concession runs the other way. CIDOC-CRM's event model is genuinely more faithful for the many-sided transitions, the ones where one polity becomes several or several become one. A partition is not really two separate edges from British India to India and to Pakistan; it is a single event with one input and two outputs, and the event node says so directly while the property graph splits it into parallel edges. But the property graph is not throwing that information away. The edges still record that both successors came from the same predecessor in the same year, and with a little additional structure a transformation can gather them back into a single n-ary event. The fidelity is deferred, not lost.

That deferral is the actual proposal. Build in a property graph because it is the tractable, legible working model; generate RDF as a publication layer when interoperability with the linked-data ecosystem matters. The European Holocaust Research Infrastructure (EHRI) is the clearest worked example: its portal store is Neo4j, while EHRI-KG (García-González and Bryant 2023) exposes the same material as RDF aligned to RiC-O and schema.org for downstream consumption. The confidence problem does not disappear in this approach, but it is bounded. Turning this project's typed edges into event-based RDF is a one-time transformation run against a stable dataset, and verifying that the result is genuinely CIDOC-CRM-compliant is a single end-of-process review rather than a discipline imposed on every act of data entry. That is a far more tractable place to need ontology expertise. And because every territory already carries a Wikidata QID, entity-level interoperability is in place now; the RDF export would add the relational layer on top of it.

E52 Time-Span and the partition/merger/federation/succession distinction carried by P2 has type. I offer this as a sketch rather than an authoritative model: CIDOC-CRM is intricate enough that being confident any given version is correct is genuinely hard, which is itself part of the argument below.

So this is not a rejection of CIDOC-CRM. It is a sequence. The property graph is where the history gets built and checked, by a historian and a coding agent, at the level of detail the sources support. RDF modelled in CIDOC-CRM is where it can go when the project, or someone building on it, needs to compose with the wider linked-data world. Keeping the working model light is what keeps that door open rather than closing it.

2.4 Schema

Every territory carries the base label `:HistoricalTerritory`, which lets pattern queries that should apply to all polities (lineage traversals, date filters, regional groupings) work uniformly. Most colonial territories also carry `:Colony` as a second, near-universal label; princely states, independent nations, and the Boer republics do not. Beyond that, each territory carries one or more more-specific subtype labels, and sixteen are in use across the graph: `:Colony`, `:CrownColony`, `:Protectorate`, `:Dominion`, `:Mandate`, `:PrincelyState`, `:Federation`, `:IndependentNation`, `:Province`, `:CompanyTerritory`, `:MinorTerritory`, `:Dependency`, `:MilitaryAdministration`, `:BoerRepublic`, `:OverseasTerritory`, and `:Condominium`. Multiple labels are permitted, so that a territory which began as a colony and became a dominion carries both, with the transition encoded in the relationship that links the earlier and later configurations. Queries can then ask either the polity-shape question ("which dominions existed in 1925?") or the lineage question ("trace this territory's status changes over time") without rewriting. The subtype set is larger than a tidy controlled vocabulary would be: it grew as the sources introduced cases the original labels did not cover, and trimming it is part of the unfinished audit described below.

Seven relationship types carry the great majority of the graph's 1,203 edges, and the distinctions between them matter historically. A partition (`:PARTITIONED_INTO`) is one entity becoming several, as when British India became India and Pakistan in 1947. A merger (`:MERGED_INTO`) is several entities becoming one, as when Upper and Lower Canada merged into the Province of Canada in 1841. A federation (`:FEDERATED_INTO`) is several entities joining a new larger entity without ceasing to exist themselves, as when the four founding provinces formed the Dominion of Canada in 1867. An evolution (`:EVOLVED_INTO`) is one entity becoming the next configuration of itself with continuous legal personality. Independence (`:BECAME_INDEPENDENT`) marks the transition from colonial to sovereign status. Administration (`:ADMINISTERED_UNDER`) captures concurrent governance arrangements where one entity is run from another without being absorbed by it. `:PART_OF` captures spatial inclusion. Together these seven account for roughly 89 percent of the edges, with `:EVOLVED_INTO` (508 edges) and `:ADMINISTERED_UNDER` (436) by far the most common.

A further fourteen relationship types appear more rarely, and they fall into two groups. Some are near-synonyms of the core seven that a consolidation pass should fold back in, with the finer distinction moved to an edge property: `:SUCCEEDED` (66 edges), `:REORGANIZED_AS` (15), `:INCORPORATED_INTO` (8), `:REUNITED_INTO` (2), and the status-change family `:BECAME_COLONY`, `:BECAME_CROWN_COLONY`, `:BECAME_PROTECTORATE`, `:BECAME_MANDATE`, and `:BECAME_SEPARATE_COLONY`. Others capture relations the core vocabulary genuinely does not express: `:TRANSFERRED_SOVEREIGNTY`

(12) and `:TRANSFERRED_TERRITORY` (1) record cession to a non-British power, `:BORDERS_WITH` (12) and `:NEAR_COAST_OF` (3) record spatial adjacency, and `:WAS_MEMBER_OF` (9) records membership.

Edges can carry properties, but most in the current graph do not: of the 1,203 relationships, 998 carry none, 181 carry a `year`, and smaller numbers carry `source`, `description`, `detail`, or `succession_type`. The capacity to attach a date and a citation to every assertion is built into the model; populating it consistently is unfinished work.

The discipline of restraint matters here, and it is worth being explicit both about why and about where this graph falls short of it. Property graphs let a curator create any relationship type at any time. There is no schema validation forcing edge types to be declared in advance, and no warning if a new one is introduced by typo. That freedom is appealing in early modelling, when the right distinctions are not yet clear, but the cost shows up later in querying. A pattern like `[:PARTITIONED_INTO|:FEDERATED_INTO|:EVOLVED_INTO*]` only returns the right answers if those relationship types are used consistently across the entire dataset. If half the partitions are encoded as `:PARTITIONED_INTO` and the other half as `:SUCCEEDED` or `:REORGANIZED_AS`, the query silently returns the wrong results, and the wrongness is invisible because no error is raised. This graph has exactly that exposure: as the inventory above shows, several of its twenty-one relationship types are near-synonyms that a disciplined consolidation would fold into the core seven. The discipline that produces a reliably queryable graph is restraint: keep the relationship-type vocabulary as small as the work allows, audit the graph periodically for typos and near-synonyms, and treat adding a new edge type as a methodological decision rather than a casual choice. Naming the gap is the point. The discipline is a practice, not a state, and an honest schema description should show where the practice is still catching up.

The honest comparison with the standards-based approach is that a formal ontology pushes a project toward declaring its schema up front. That feels heavy and constrains exploration, but it catches inconsistency mechanically. Property graphs feel light and let a model evolve as the source material is better understood, but the curator becomes the schema's only enforcer, and this graph shows what that costs when the enforcement lags. Both choices are defensible. Historians adopting the property-graph model for their own work should know they are signing up for the second deal.

3 Future work

This paper is the first in a planned series. A second paper will document the use of the Wikidata MCP server to ground entities at scale, including the disambiguation patterns that worked, the ones that didn't, and the verification workflow that kept hallucinated QIDs out of the graph. A third, longer paper will extend the territorial scaffolding presented here into a much larger knowledge graph built from the *Colonial Office List* for the period 1867 to 1968, populating the colonies with the personnel, offices, appointments, and administrative arrangements that actually ran them. Each paper builds on the convention argued for above: ground to Wikidata, keep your own model,

publish so others can join. The graph in this paper is intended as a substrate the later work can attach to, and as an invitation to anyone whose corpus would compose with it.

The dataset is also a natural candidate for extension to the other European empires of the same period: French, Spanish, Portuguese, Dutch, Belgian, German, and Italian. Beyond Europe, the same modelling approach could be extended to non-European political entities whose territorial histories are entangled with the imperial frame the British graph currently centres. Each of those extensions would benefit from the same Wikidata-grounded approach, and each would compose with the present graph automatically wherever the QID layer is shared.

3.1 A note on what a graph of imperial territories cannot represent

A serious limitation concerns the kinds of sovereignty a graph of empires represents and the kinds it cannot. Modelling Indigenous nations in linked open data bumps against the OCAP principles of Indigenous data ownership, control, access, and possession, and global-history projects need to be careful about what they assert in a knowledge graph, particularly around sovereignty. The graph in this paper records British imperial claims to territory; it does not record who actually exercised authority on the ground. Rupert's Land, for example, is included as a node because the Hudson's Bay Company held a legal charter to it, but the company did not exercise the wide territorial control or monopoly of violence characteristic of a settler colony, and large parts of the chartered area remained under Indigenous governance throughout the HBC period. Likewise, the formal expansion of Canada in the 1870s to incorporate British Columbia and the North-West Territories is encoded in the graph as a federation and a series of accessions, but the inclusion of these transitions is not an endorsement of Canadian claims to unceded lands, nor of treaty processes whose Indigenous signatories did not understand themselves to be surrendering sovereignty. The graph is a record of one set of claims and should be read alongside scholarship that takes Indigenous sovereignty as its starting point.

4 Code and data availability

The dataset, visualization, and loading scripts are openly available at https://github.com/Working-Papers-in-Critical-Search/paper-002-empire-evolution. The repository contains:

- `data/britishempire_kg_export.cypher`: the full graph: 747 historical territories (314 colonial polities and 433 princely states) and 1,203 typed relationships, exported from Neo4j.
- `data/qid_manifest.tsv`: every territory with its Wikidata QID and, where applicable, the scope note documenting which stand-in entity was used and why.
- `viz/empire_evolution.html`: the self-contained D3.js Sankey visualization shown in the figures above. The data is embedded in the HTML; open the file in any browser.
- `scripts/load_falkordb.py`: an in-process loader for FalkorDB (no Neo4j server, no Docker, no Java required).

- `notebooks/quick_tour.ipynb`: a guided tour with three pyvis subgraph visualizations (schema-level statistics, the Canada lineage tree, and the Southeast Asia regional subgraph).

4.1 Loading the graph

The lightest-weight option is the embedded FalkorDB loader, which spins up a graph in-process and drops the user into a Cypher REPL:

```
pip install -r requirements.txt
python3 scripts/load_falkordb.py --interactive
```

For a full Neo4j installation, the same export loads with `cypher-shell`:

```
cypher-shell -u neo4j -p <password> -f data/britishempire_kg_export.cypher
```

Neo4j 5.x is recommended. Re-running the script is idempotent.

4.2 Sample queries

Territories created by partition, with the year each successor was established:

```
MATCH (a:HistoricalTerritory)-[:PARTITIONED_INTO]->(b:HistoricalTerritory)
RETURN a.canonical_name, b.canonical_name, b.established_year
ORDER BY b.established_year;
```

Lineage of any territory back to its earliest predecessor:

```
MATCH path = (root:HistoricalTerritory)-[:EVOLVED_INTO|:PARTITIONED_INTO|:MERGED_INTO*]->
  (target:HistoricalTerritory {canonical_name: 'Canada, Dominion of'})
WHERE NOT (()-[:EVOLVED_INTO|:PARTITIONED_INTO|:MERGED_INTO]->(root))
RETURN path;
```

References

Cha, Javier. 2026. "Fine-Tuning the Historian's Macroscope: Data Reuse and Medieval Korean Biographical Records in Neo4j." *Journal of Cultural Analytics* 11 (1). https://doi.org/10.22148/jca.1027.

Edge, Darren, Ha Trinh, Newman Cheng, et al. 2024. *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. https://arxiv.org/abs/2404.16130.

García-González, Herminio, and Mike Bryant. 2023. "The Holocaust Archival Material Knowledge Graph." In *The Semantic Web – ISWC 2023*, vol. 14266. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-031-47243-5_20.

Han, Haoyu, Yu Wang, Harry Shomer, et al. 2025. *Retrieval-Augmented Generation with Graphs (GraphRAG)*. https://arxiv.org/abs/2501.00309.

MacKay, Jon. 2018. *Dealing with Big Data and Network Analysis Using Neo4j*. Programming Historian. https://doi.org/10.46430/phen0074.

Pan, Keyao. 2022. "Networking for Historical Justice: The Application of Graph Database Management Systems to Network Analysis Projects and the Case Study of the Reparation Movement for Japanese Colonial and Wartime Atrocities." *Journal of Open Humanities Data* 8: 11. https://doi.org/10.5334/johd.76.

Varga, Rada, and Stefan Bornhofen. 2024. "Graph Based Modelling of Prosopographical Datasets. Case Study: Romans 1by1." *Digital Humanities Quarterly* 18 (2). https://doi.org/10.63744/sm8dnje5gx35.

Warren, Christopher N., Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. 2016. "Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks." *Digital Humanities Quarterly* 10 (3). https://doi.org/10.63744/2dzv4awt8pgh.