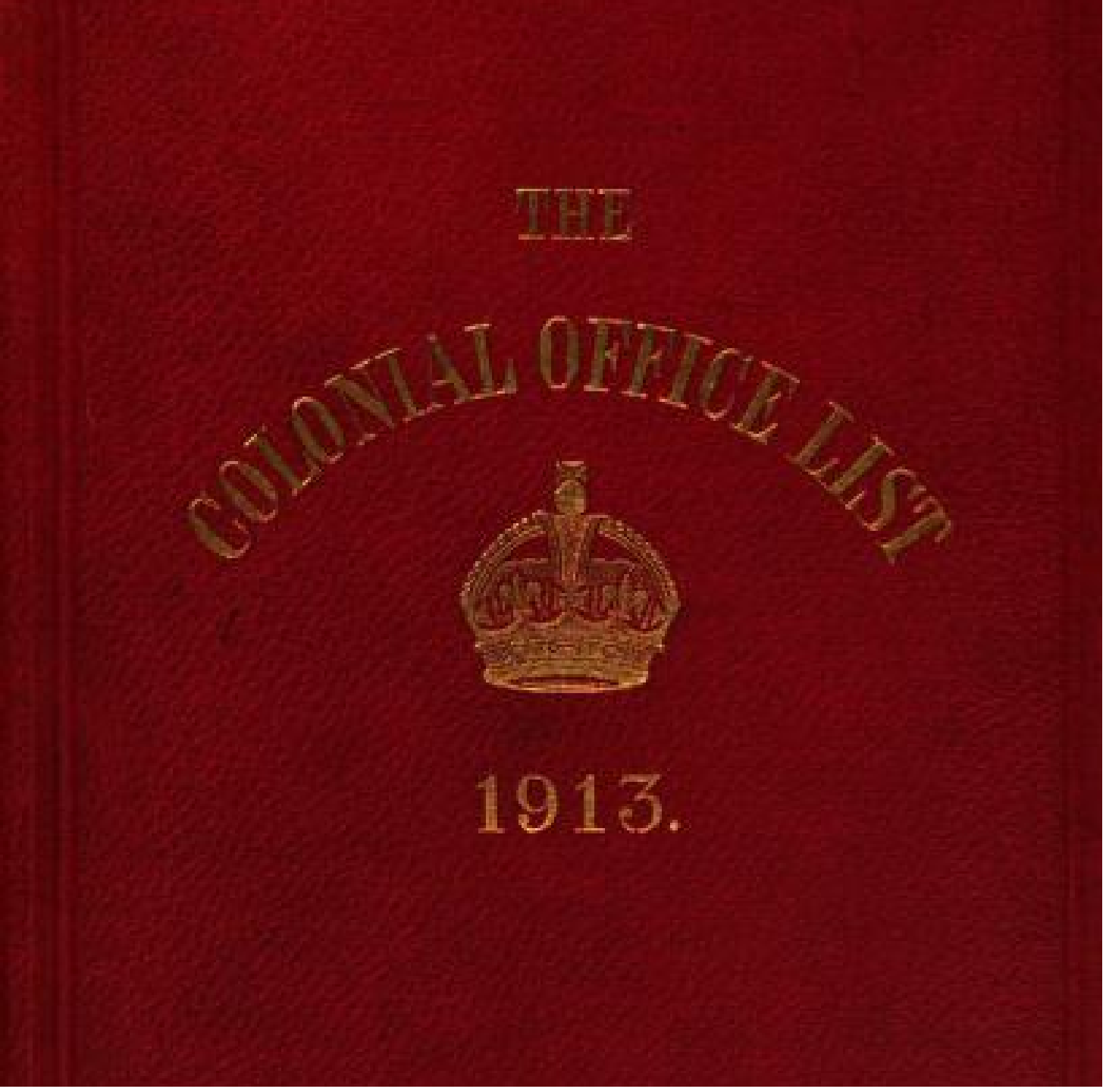


Text as Historical Data in 2026

JIM CLIFFORD

ASSOCIATE PROFESSOR

UNIVERSITY OF SASKATCHEWAN

Internet Archive

01

At least 500 million pages of out of copyright text

02

Open access

03

Uneven metadata

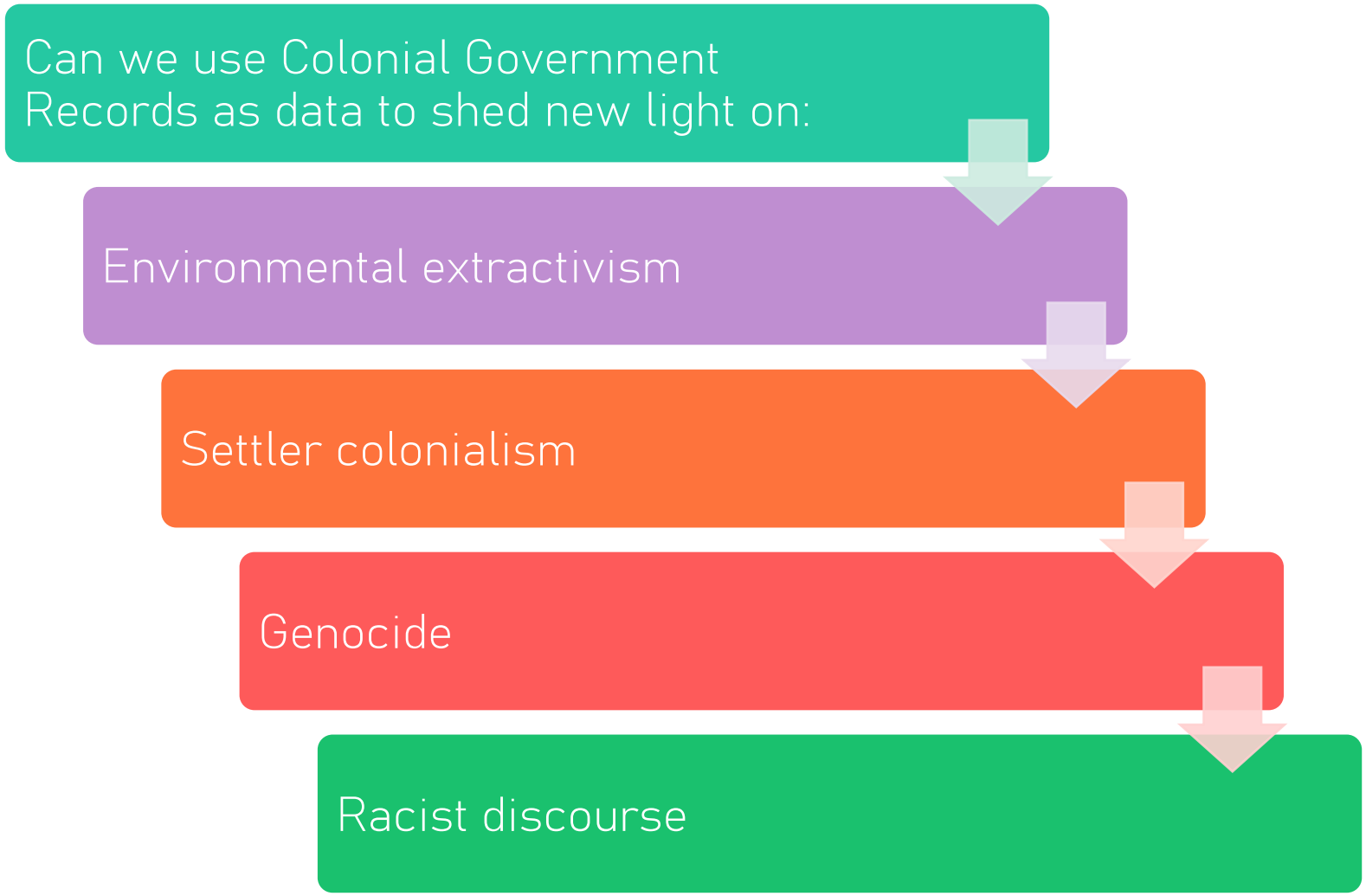
04

English and western European languages overrepresented

05

Dearth of subaltern voices

Official Government Documents



Colonial Office List

- Mix between a staff directory, encyclopedia and short yearbook for most British Colonies (the Indian Office had its own parallel volumes)
- Broken down by colonies
- Includes:
  - Basic history
  - Economic and trade information
  - Government structures
  - List of government employees with salary, rank, location

Semi-structured data

01

The volumes follow a pattern, but they also have lots of exceptions to the rules

02

olmOCR creates relatively clean Markdown text

03

How do we extract all the people, with their attribute data, into a database and link people across years?

Linked Open Data

How do we ground place names (toponyms) and people to existing linked open data precedent identifiers like Geonames and Wikidata ids?

How do we disambiguate John Smith?

Persistent Identifiers for People in the Past

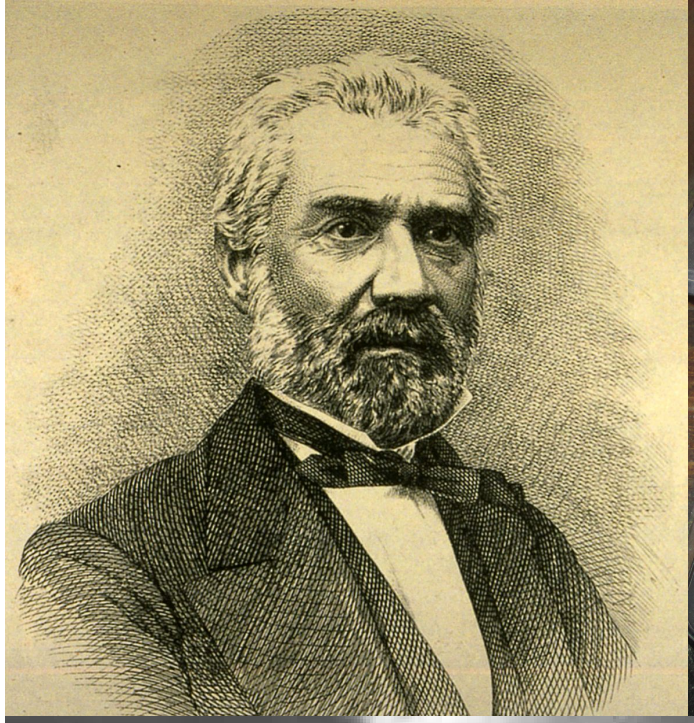
JIM CLIFFORD AND MATTHEW KUNKEL, SASKATCHEWAN

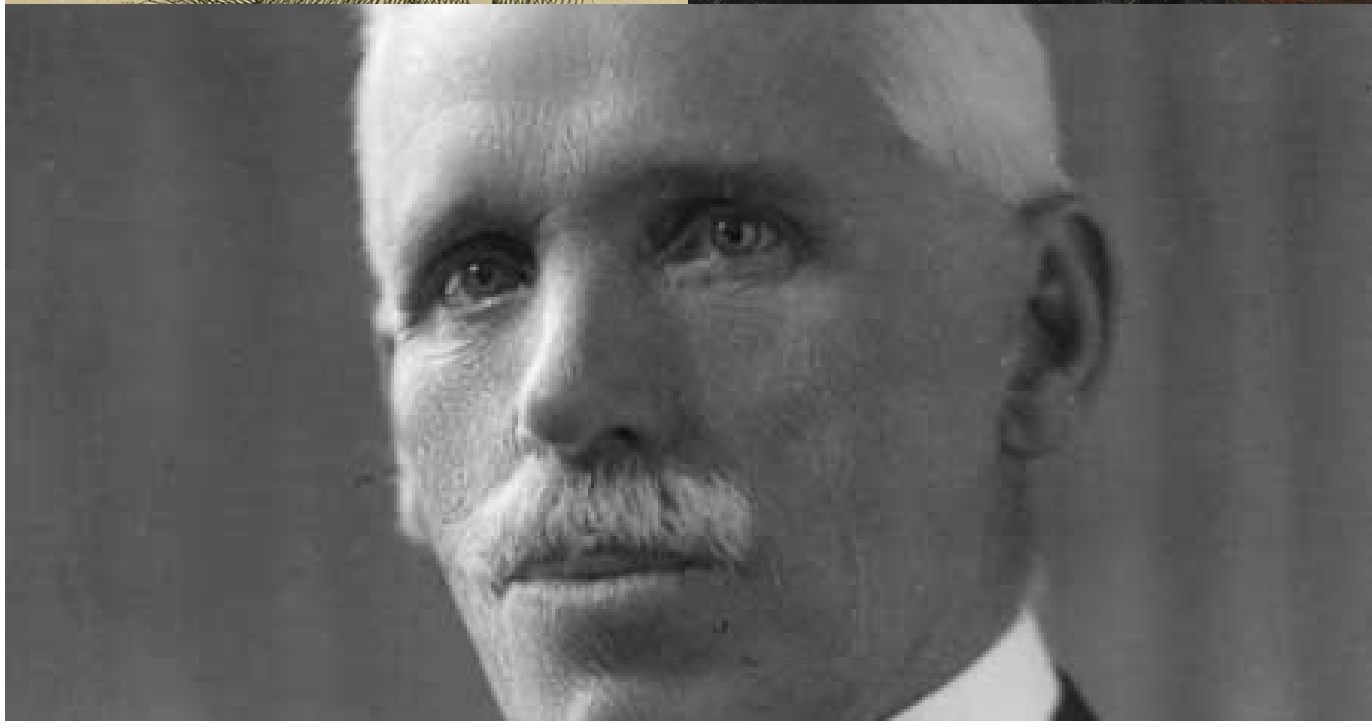
SUSAN BROWN AND SARAH ROGER, GUELPH

NATALIE HERVIEUX, ALBERTA

Three James Armstrongs

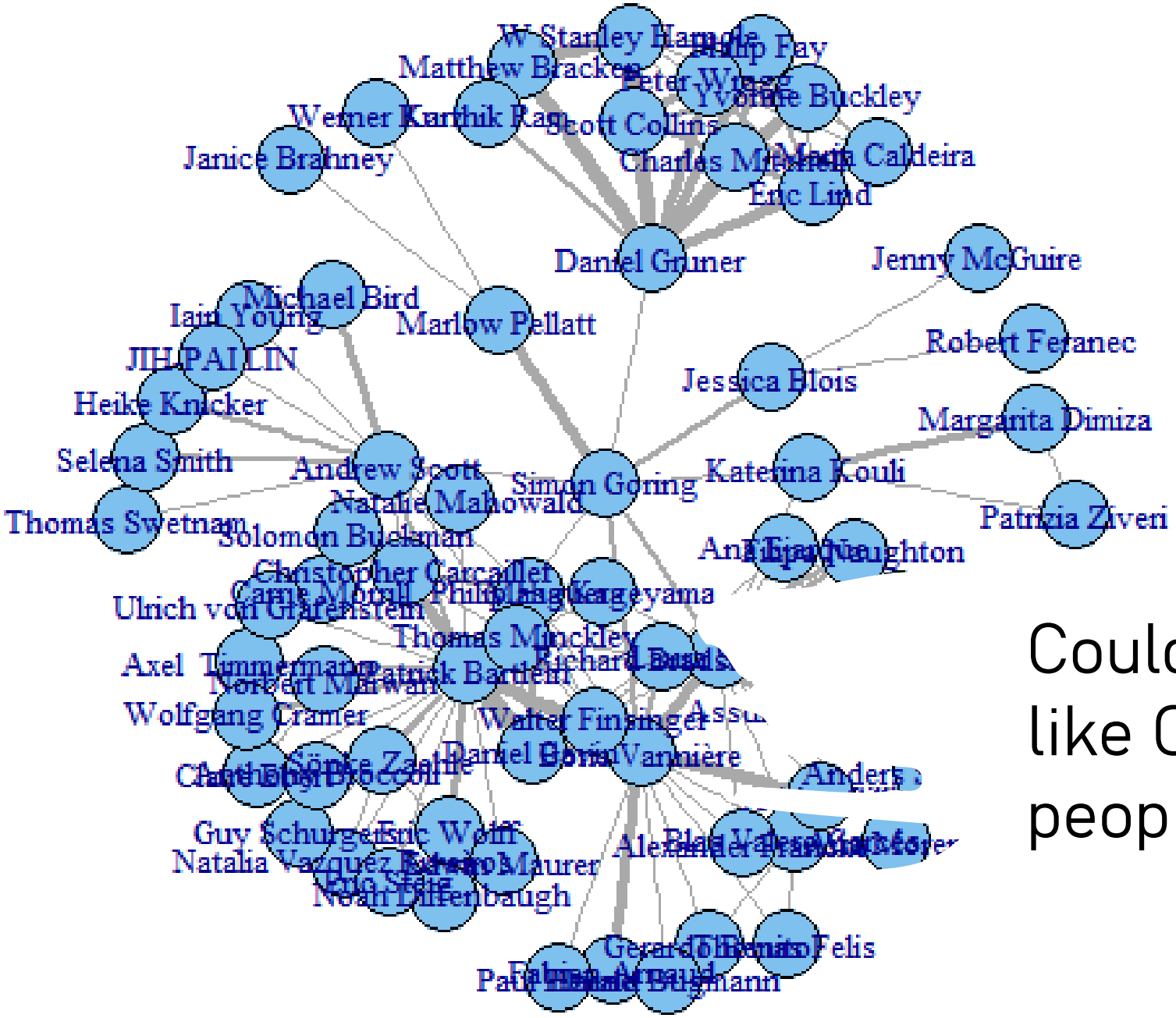
- Which James Armstrong worked for the Department of Indian Affairs?
- It probably wasn't the guy who made the stoves.

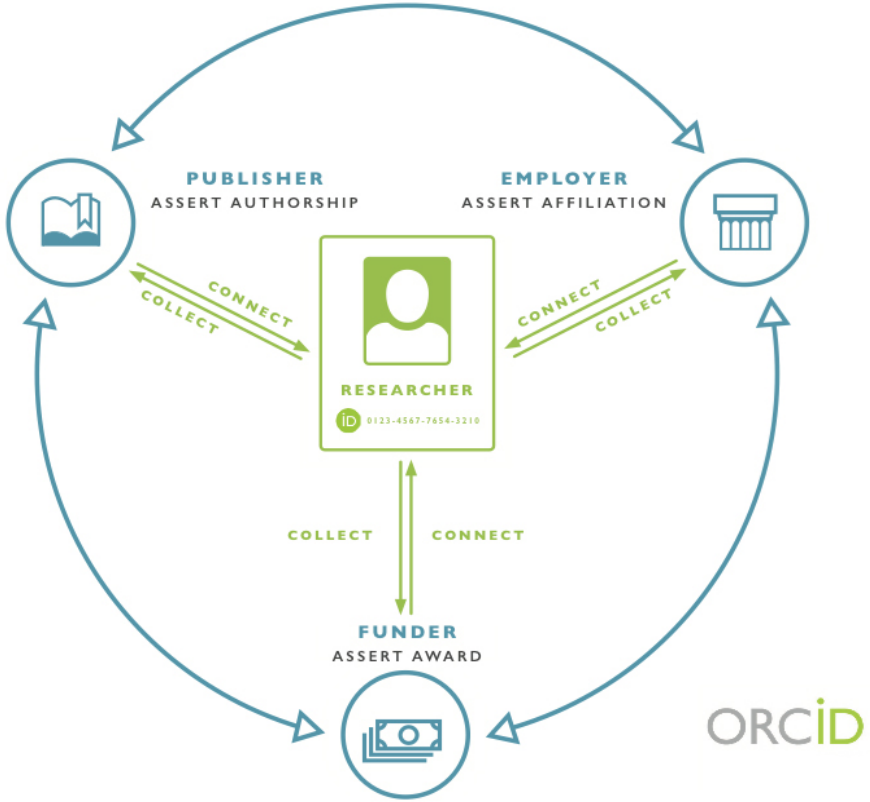




One person, Multiple Names

- **Esther Brandeau or Jacques La Fargue**
- Thayendanegea, Thayendanegen, Thayeadanegea, Joseph Thayendanegea, or Joseph Brant
- Félicité Angers or Laure Conan
- Hilda Doris Buck or Doris Hilda Anderson

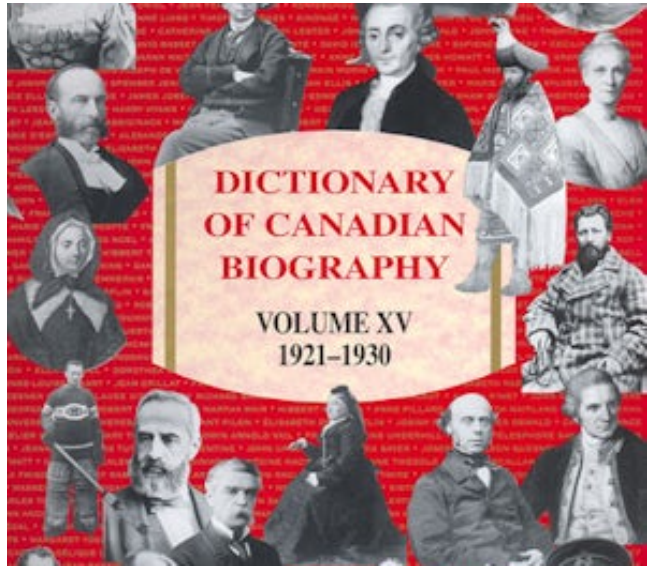


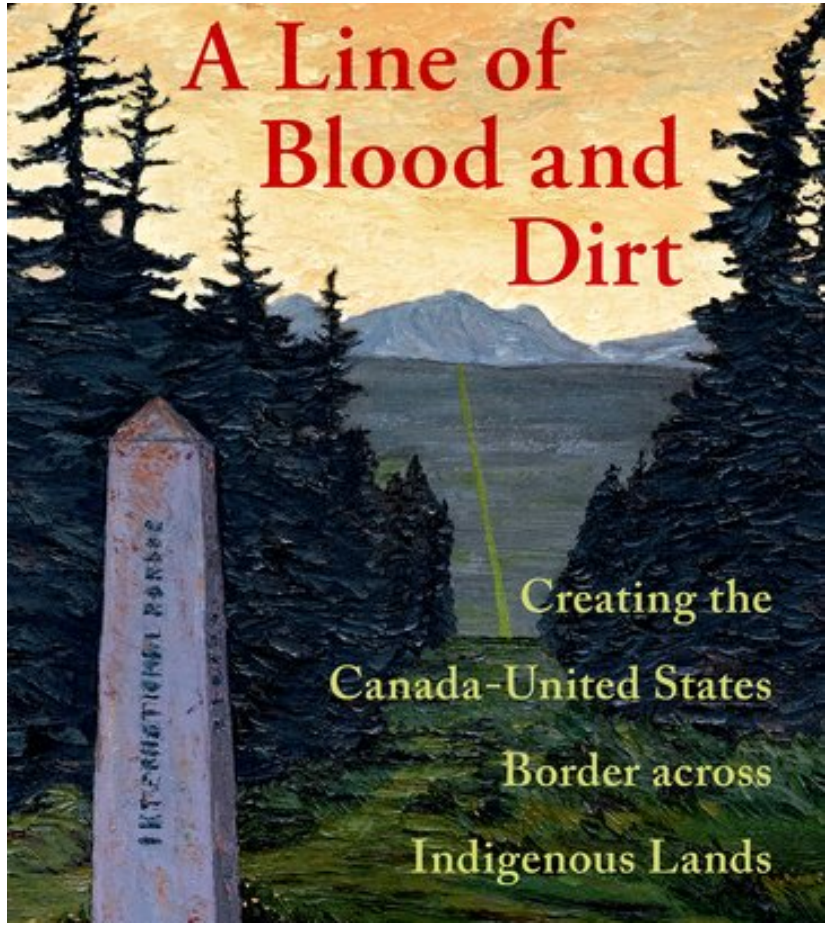


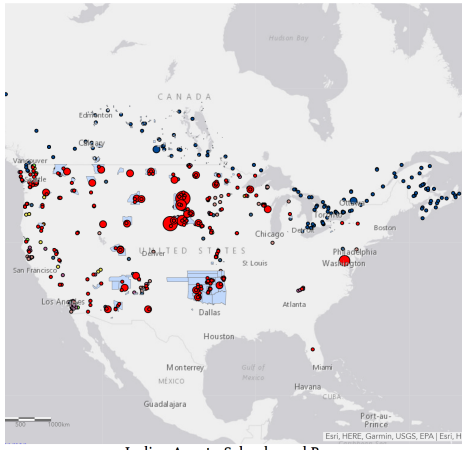
Could we create something like ORCID for historical people?

Where to start?

- Dictionary of Canadian Biography
- The project started in 1959
- 8000+ biographies published in French and English
- Focused on elite men of European decent but working hard to correct that legacy
- Planning to launch a new version of the website in 2024







Indian Agents, Schools, and Reserves

Ben Hoy's Department of Indian Affairs Employee Data

- Hoy created a database of Department of Indian Affairs employees' locations in the late 19th and early 20th centuries.
- The database includes **limited** biographical information: names, occupation, location and years employed by the DIA.
- The individuals and their correspondence are an important part of the historical record of settler colonialism in Canada.
- Opportunity to test creating PIDs for people without full biographies.

Key objectives and principals

1

Start by reconciling data against existing PIDs

2

Use names, dates, places, and occupations to disambiguate individuals

3

Avoid creating certainty from uncertainty when parsing text into data points

4

Avoid duplicating the use of offensive terms in dated reference materials

Modeling

- Started simple, but quickly grew.
  - Added links to both French and English biographies
  - Added groups, including Department of Indian Affairs
  - Added links to archival fonds
  - Added layers to occupations
  - Added models to handle uncertainty
- Big thanks to Jessica Ye

Identifiable Properties

- Names
- Occupations
- Birth date/place
- Family connections
- Death date/place

Utilize existing DCB structures

Existing Categorizations →

- Browse by Region of Birth
- Browse by Region of Activities
- Browse by Occupations and Other Identifiers
- Browse by Name of Subject A-Z

Leverage Formulaic Structure →

BOURASSA, FRANÇOIS, farmer, Patriote, militia officer, and politician; b. 5 June 1813 in Sainte-Marguerite-de-Blairfindie (L'Acadie), Lower Canada, eldest son of François Bourassa, its first mayor, and Geneviève Patenaude; brother of Napoléon Bourassa* and uncle of Henri Bourassa*; d. 13 May 1898 in Saint-Valentin, Que.

DCB Name Extraction

McKENZIE, JAMES, fur trader, JP, and businessman; b. *c.* 1777 near Inverness, Scotland, son of Alexander Mackenzie and his wife Catherine; d. 18 July 1849 at Quebec.

BAUDRY (Beaudry), *dit* **Saint-Martin (Baudry Desbuttes, Baudry Soulard), JEAN-BAPTISTE**, gunsmith; baptized 3 July 1684 at Trois-Rivières, son of Guillaume Baudry*, *dit* Des Butes, gunsmith and silversmith, and Jeanne Soullard; m. on 8 Oct. 1721 Marie-Louise, daughter of the armourer Nicolas Doyon, in Quebec; d. 20 Nov. 1755 at Detroit.

In using numerous surnames Jean-Baptiste Baudry, it seems, wanted to distinguish himself from a namesake who lived at the same period and who also practised the armourer's craft. Through his father, mother, and wife he belonged

---

OGIMAUH-BINAESSIH (Okemapenesse, meaning "chief little bird"; **Wageezhegome, Wakeshogomy, Weggishgomin**, meaning "who is like the day"; **John Cameron, Captain John**), Mississauga Ojibwa chief, member of the eagle clan, and farmer; b. May 1764 at the Credit River (Ont.); d. 30 Sept. 1828 at the Credit Mission (Mississauga), Upper Canada.

Uncertainty options

- Leave out uncertain information
- Include uncertain information (and deal with it later)
- Add additional modelling to denote that something is uncertain

Birth Excerpt Date and Location	Birth Date	Birth Date_Exact_Sta	Birth Date_Exact_E	Birth Date_Uncert
b. about 1855 in Baffin Island (Nunavut), possibl	about 1855	1855	1855	Uncertain
b. apparently in the 1670s in Portugal, probably	apparently in the 1670s	1670-01-01	1679-12-31	Uncertain
b. c. 1805 or c. 1808, probably in Brooklyn, N.Y.	c. 1805 or c. 1808	1805-01-01	1808-12-31	Uncertain
b. in all likelihood in 1789 in Italy, probably in Ro	1789	1789	1789	Uncertain
b. most probably in January 1799 in Trois-Rivièr	January 1799	1799-01-01	1799-01-31	Uncertain
b. possibly c. 1770, probably in a Shawnee villa	possibly c. 1770	1770-01-01	1770-12-31	Uncertain
b. possibly in Bay de Verde, Nfld., about 1680	about 1680	1680	1680	Uncertain
b. probably during the late 1830s, likely in the R	probably during the late 1830s	1836	1839	Uncertain
b. probably in 1683, perhaps in London, Englan	probably in 1683	1683	1683	Uncertain
b. probably in 1757, possibly in Kilgraston, Scotl	probably in 1757	1757	1757	Uncertain
b. probably in 1768, perhaps at Arbre Croche (H	probably in 1768	1768	1768	Uncertain
b. probably in Écully, France, around 1646–48	around 1646	1646	1648	Uncertain
b. probably in France about 1625	about 1625	1625	1625	Uncertain

Name(s) / Label(s)

Date of birth of Louis O'Soup

Type(s)

medium certainty

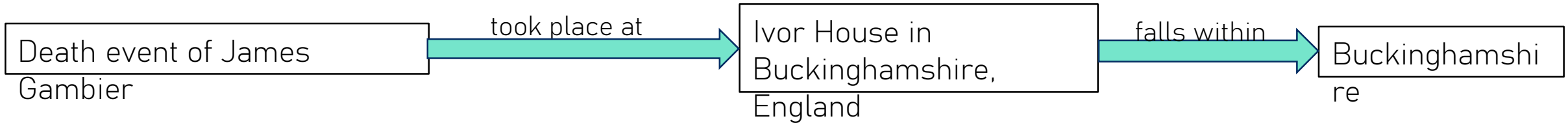
medium precision

Date Begins	1836-1-1
Date Ends	1839-12-31
Date Occured Within	probably during the late 1830s

carried out by	Dictionary of Canadian Biography

Location Reconciliation

Ireland, apparently in County Tipperary	Uncertain place - went up a level	https://www.wikidata.org/wiki/Q27
Ireland, possibly in County Kilkenny	Uncertain place - went up a level	https://www.wikidata.org/wiki/Q27
Ireland, probably in Clonmel (Republic of Ireland)	Uncertain place - went up a level	https://www.wikidata.org/wiki/Q27
Ivedon Penn, near Honiton, England	Place too specific - went up a level	https://www.geonames.org/2646658/honiton.html
Ivor House in Buckinghamshire, England	Place too specific - went up a level	https://www.wikidata.org/wiki/Q67285329
Juniper Bank, near Walkerburn, Scotland	Place too specific - went up a level	https://www.geonames.org/2634885/walkerburn.html



Occupation Reconciliation

PREFERRED TERM	**military occupation**
TYPE	crm:E55_Type
BROADER CONCEPT	occupation
NARROWER CONCEPTS	airman coastguard
ALTERNATIVE LABELS	artillery gunner artillery officer captain of militia colonel deputy lieutenant lieutenant lieutenant general major general military military communications military driver military escort military leadership military officer militia captain [show all 19 values]
URI	http://id.lincsproject.ca/occupation/militaryOccupation

officer in the colonial regular troops (minted URI)

↓ has type

military officer (Wikidata)

↓ has type

military occupation (LINCS)

colonel in the Connecticut militia **(minted URI)**

↓ has type

colonel (Wikidata)

↓ has type

military occupation (LINCS)

PIDs for unnamed people

Philosophical Transactions GraphRAG

- Search 8,128 Royal Society articles (1665-1869) with AI-powered answers
- Graph database structure that is built around years and dates.
  - 94K nodes, 1.1M edges
  - Future versions will expand to include more primary sources and a more sophisticated graph database.
- https://cljim22--phil-trans-graphrag-frontend.modal.run/