

# BUILDING LINKS WITH NATURAL LANGUAGE PROCESSING

2022 Smithsonian Digitisation Conference  
John Stack, Digital Director  
24 March 2022


**SCIENCE  
MUSEUM  
GROUP**



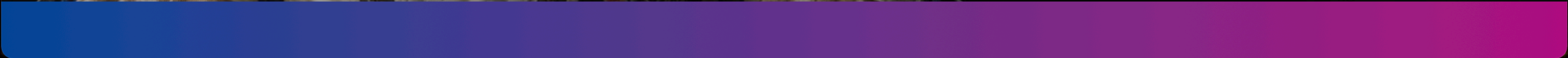




# SEARCH OUR COLLECTION

Search objects, people, categories, object numbers etc. 

Explore over 350,000 objects and archives from the Science Museum, Science and Industry Museum, National Science and Media Museum, National Railway Museum and Locomotion.





All 50

People 0

Objects 50

Documents 0

View:



Filter search



**Geoffrey Tippet Collection - Swindon Works & GWR Locomotives [supplied title]**  
Photographic Collections (Railway)  
c. 1930



**Advertisement by GWR for the exhibition of the Gal**  
Art  
1845



**Milk van, Great Western Railway**  
Locomotives and Rolling Stock  
1936



**Diesel Railcar No 4, 1934, Great Western Railway**  
Locomotives and Rolling Stock  
1934





# Zeiss Compound Microscope

MADE: [1907](#)





DETAILS

CATEGORY:	Scientific Instruments & Research
OBJECT NUMBER:	Y1991.49.2/1
TYPE:	compound microscope
TAXONOMY:	<div>furnishing and equipment<ul style="list-style-type: none"><li>tools &amp; equipment<ul style="list-style-type: none"><li>optical instrument<ul style="list-style-type: none"><li>microscope</li></ul></li></ul></li><li>furnishing and equipment<ul style="list-style-type: none"><li>tools &amp; equipment<ul style="list-style-type: none"><li>optical instrument<ul style="list-style-type: none"><li>microscope</li></ul></li></ul></li></ul></li></ul></div>
CREDIT:	Gift of Central Manchester Health Authority

CITE THIS PAGE

Science Museum Group. Zeiss Compound

RIGHTS

We encourage the use and reuse of our collection data.

Data in the title, made, maker and details fields are released under Creative Commons Zero

Descriptions and all other text content are licensed under a Creative Commons Attribution 4.0 licence

Using our data

DOWNLOAD

Download catalogue entry as JSON

View manifest in IIIF viewer

Add to Animal Crossing Art Generator

Download manifest IIIF

Our records are constantly being enhanced and improved, but please note that we cannot guarantee the accuracy of any information shown on this website.



# DISCOVERY AND EXPLORATION

1. Knowing about the collection
2. Understanding that it may contain relevant contents
3. Finding that content is available online
4. Navigating keyword-based search
5. Grasp the relevance of the content to the search
6. Navigate from that content to related content



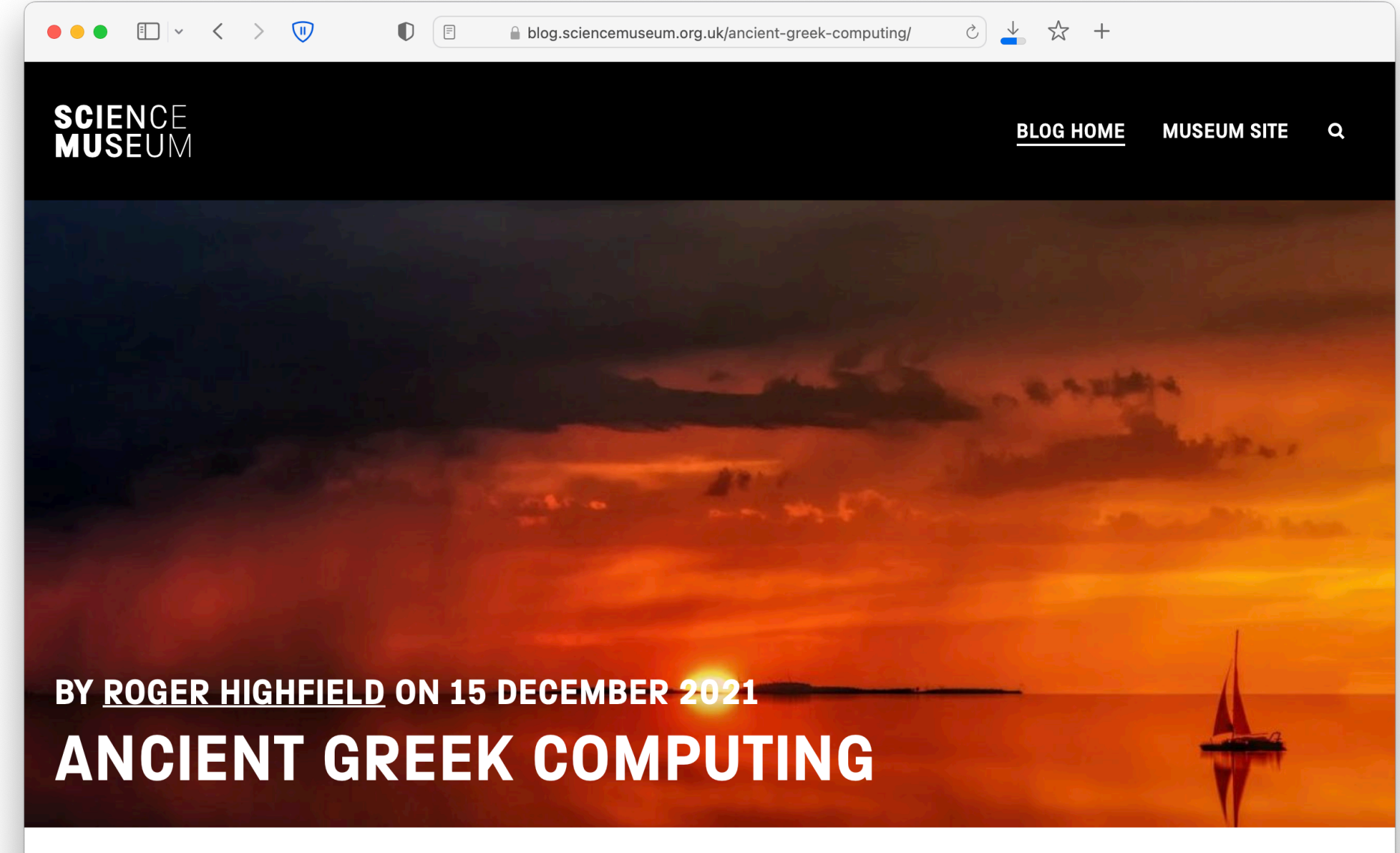
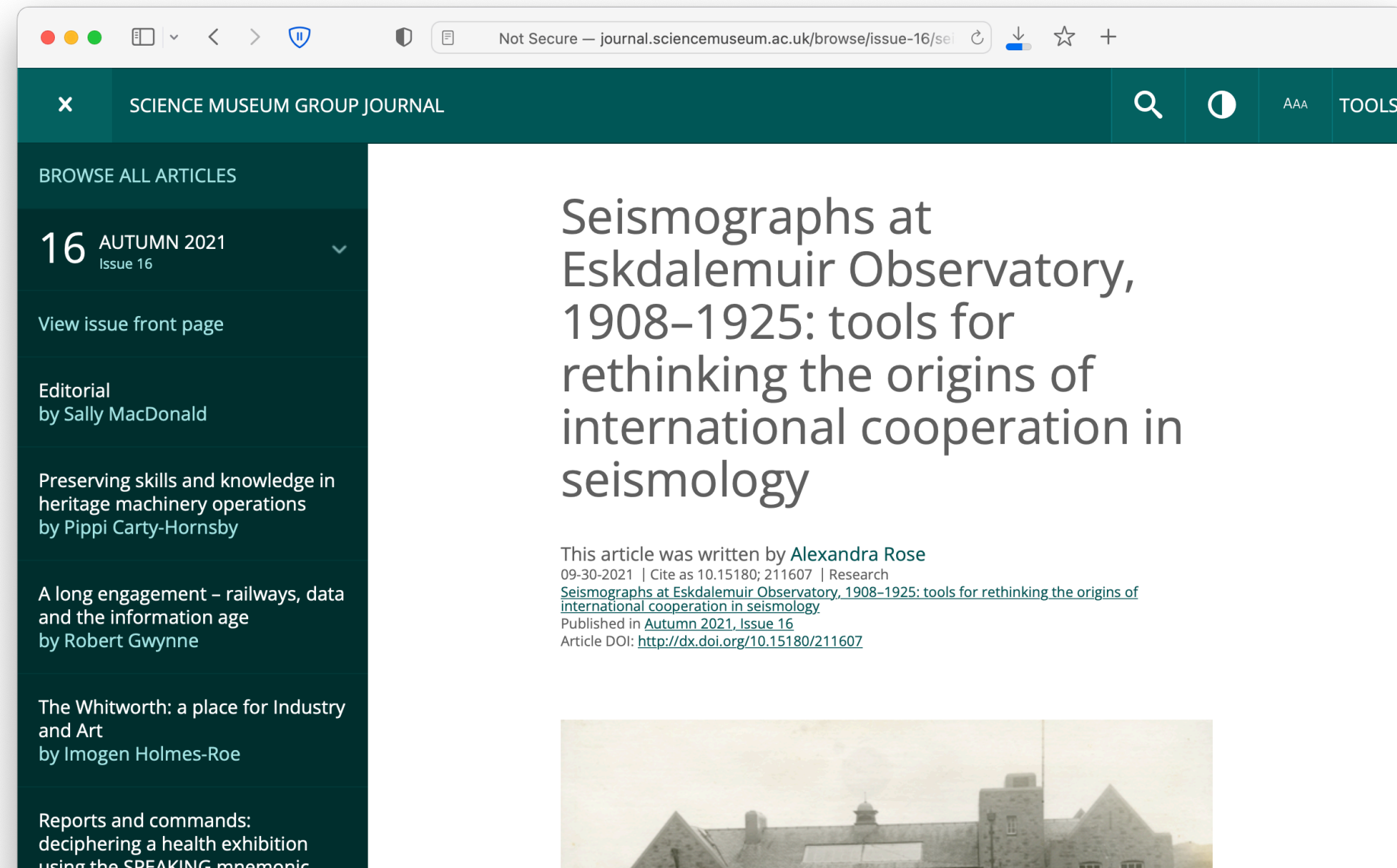
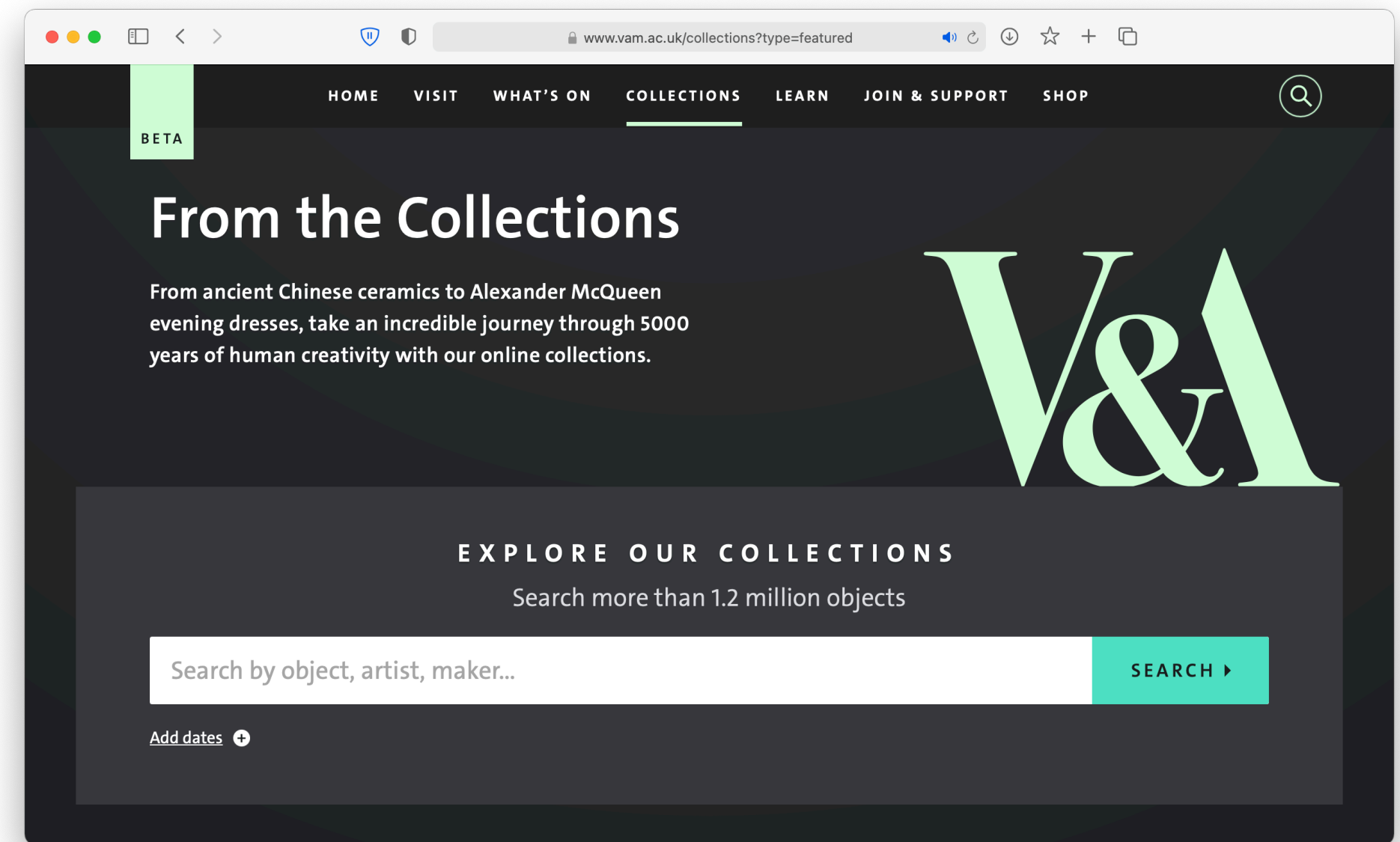
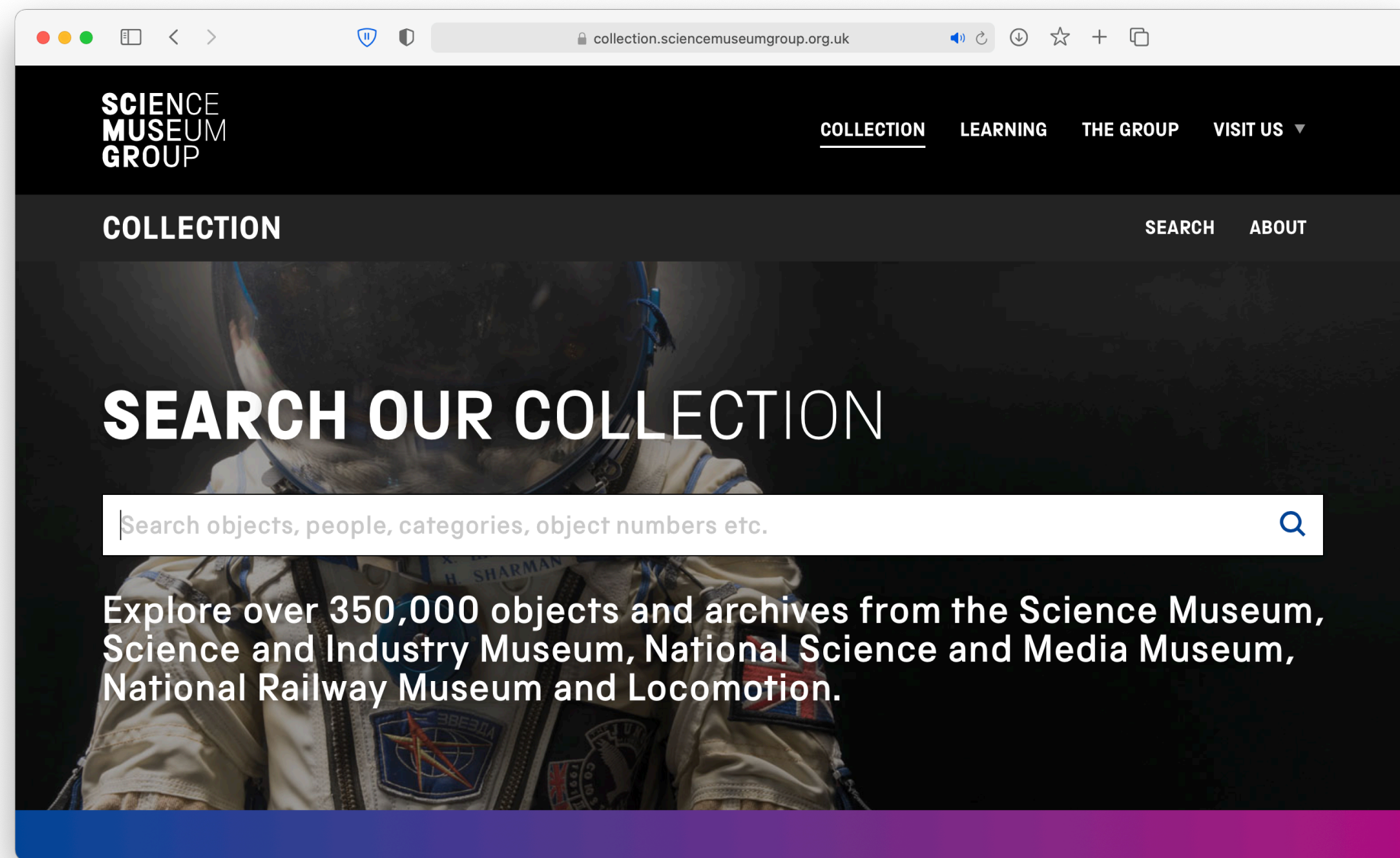
# HERITAGE CONNECTOR PROJECT

How can existing digital tools and methods be used to build relationships at scale between poorly and inconsistently catalogued digitised collection objects and other content sources?



# FIVE CONTENT SOURCES













- Main page
- Community portal
- Project chat
- Create a new Item
- Recent changes
- Random Item
- Query Service
- Nearby
- Help
- Donate

- Lexicographical data
- Create a new Lexeme
- Recent changes
- Random Lexeme

- Tools
- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Cite this page
- Concept URI

Item Discussion

Read View history

Search Wikidata

# General Electric (Q54173)

American multinational conglomerate corporation [edit](#)  
General Electric Company | GE | General Electric Co.

In more languages  
Configure

Language	Label	Description	Also known as
English	General Electric	American multinational conglomerate corporation	General Electric Company GE General Electric Co.
British English	General Electric	No description defined	
French	General Electric	conglomérat américain	
German	General Electric	US-amerikanischer Mischkonzern	

All entered languages

## Statements

instance of	<div><div></div><div>business</div><div><a href="#">edit</a></div></div> <div><div></div><div>0 references</div><div><a href="#">+ add reference</a></div></div>
	<div><div></div><div>enterprise</div><div><a href="#">edit</a></div></div> <div><div></div><div>0 references</div><div></div></div>



🏠

Marks Tey Station With Mr L Sydney Bracher

MADE: 1899-06-06 in [Marks Tey Railway Station](#)  
MAKER: [Edgar Tarry Adams](#)



A photograph entitled 'Marks Tey Station With Mr L. Sydney Bracher', taken by Edgar Tarry Adams in June 1899. Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

A photograph entitled 'Marks Tey Station With Mr L Sydney Bracher', taken by Edgar Tarry Adams in June 1899.

Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

DETAILS

CATEGORY:	<a href="#">Photographs</a>
OBJECT NUMBER:	1997-5003/157
TYPE:	<a href="#">photograph</a>
TAXONOMY:	visual and verbal communication

CITE THIS PAGE

Science Museum Group. Marks Tey Statio

🔗

RIGHTS

We encourage the use and reuse of our collection data.



Data in the title, made, maker and details fields are released under [Creative Commons](#)



🏠

Marks Tey Station With Mr L Sydney Bracher

MADE: 1899-06-06 in [Marks Tey Railway Station](#)  
MAKER: [Edgar Tarry Adams](#)



'Marks Tey Station With Mr. L. Sydney Bracher, 1899 A photograph of Mr L Science Museum Group Collection © The Board of Trustees of the Science Museum

↶ ↷ > Use this image

A photograph entitled 'Marks Tey Station With Mr L. Sydney Bracher', taken by Edgar Tarry Adams in June 1899. Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

A photograph entitled 'Marks Tey Station With Mr L Sydney Bracher', taken by Edgar Tarry Adams in June 1899.

Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

DETAILS

CATEGORY:	<a href="#">Photographs</a>
OBJECT NUMBER:	1997-5003/157
TYPE:	<a href="#">photograph</a>
TAXONOMY:	visual and verbal communication

CITE THIS PAGE

Science Museum Group. Marks Tey Statio

RIGHTS

We encourage the use and reuse of our collection data.

the title, made, maker and details fields are released under [Creative Commons](#)



🏠

Marks Tey Station With Mr L Sydney Bracher

MADE: 1899-06-06 in [Marks Tey Railway Station](#)  
MAKER: [Edgar Tarry Adams](#)



A photograph entitled 'Marks Tey Station With Mr L. Sydney Bracher', taken by Edgar Tarry Adams in June 1899. Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

A photograph entitled 'Marks Tey Station With Mr L Sydney Bracher', taken by Edgar Tarry Adams in June 1899.

Mr Bracher, probably a friend of Adams, is captured smoking his pipe and tipping his straw boater at the camera at the station in Essex.

DETAILS

CATEGORY:	<a href="#">Photographs</a>
OBJECT NUMBER:	1997-5003/157
TYPE:	<a href="#">photograph</a>
TAXONOMY:	visual and verbal communication

CITE THIS PAGE

Science Museum Group. Marks Tey Statio

RIGHTS

We encourage the use and reuse of our collection data.



Data in the title, made, maker and details fields are released under [Creative Commons](#)



- Main page
- Community portal
- Project chat
- Create a new Item
- Create a new Lexeme
- Recent changes
- Random Item
- Query Service
- Nearby
- Help
- Donate
- Print/export
- Create a book
- Download as PDF
- Printable version
- Tools
- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Concept URI
- Cite this page

Marks Tey railway station (Q2025932)

No description defined

In more languages

Statements

instance of


railway station

0 references

add reference

add value

image



Markstey1.jpg

648 × 429; 98 KB

1 reference

add value

named after

Marks Tey

0 references

add reference

add value

country

United Kingdom

0 references

add reference

add value


located in the administrative territorial entity

Marks Tey

1 reference

add value

coordinate location



51°52'51"N, 0°46'58"E

1 reference

add value

operator

Greater Anglia

edit

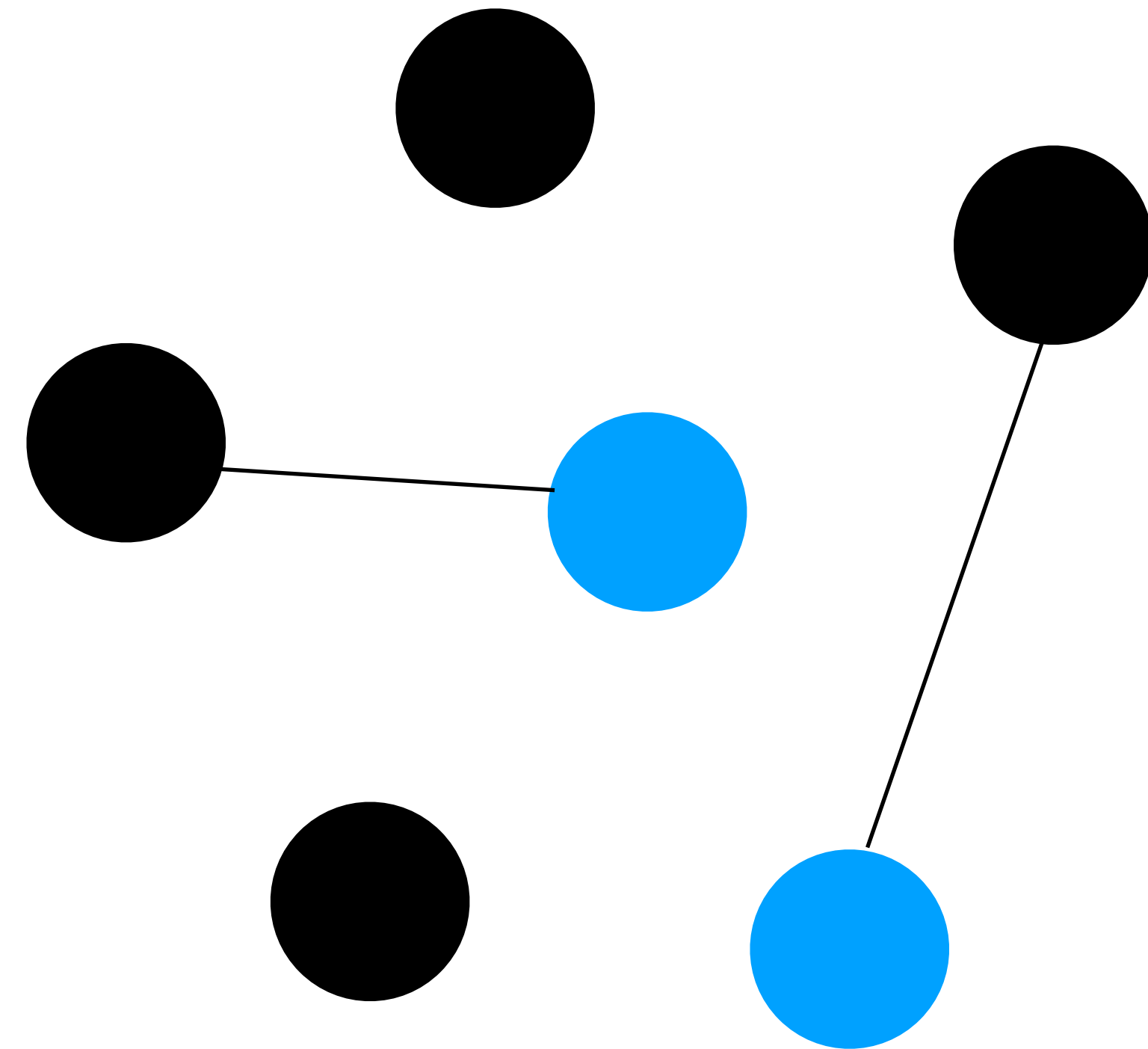


# LINK BUILDING



**THIS IS OUR  
COLLECTION  
NOW..**

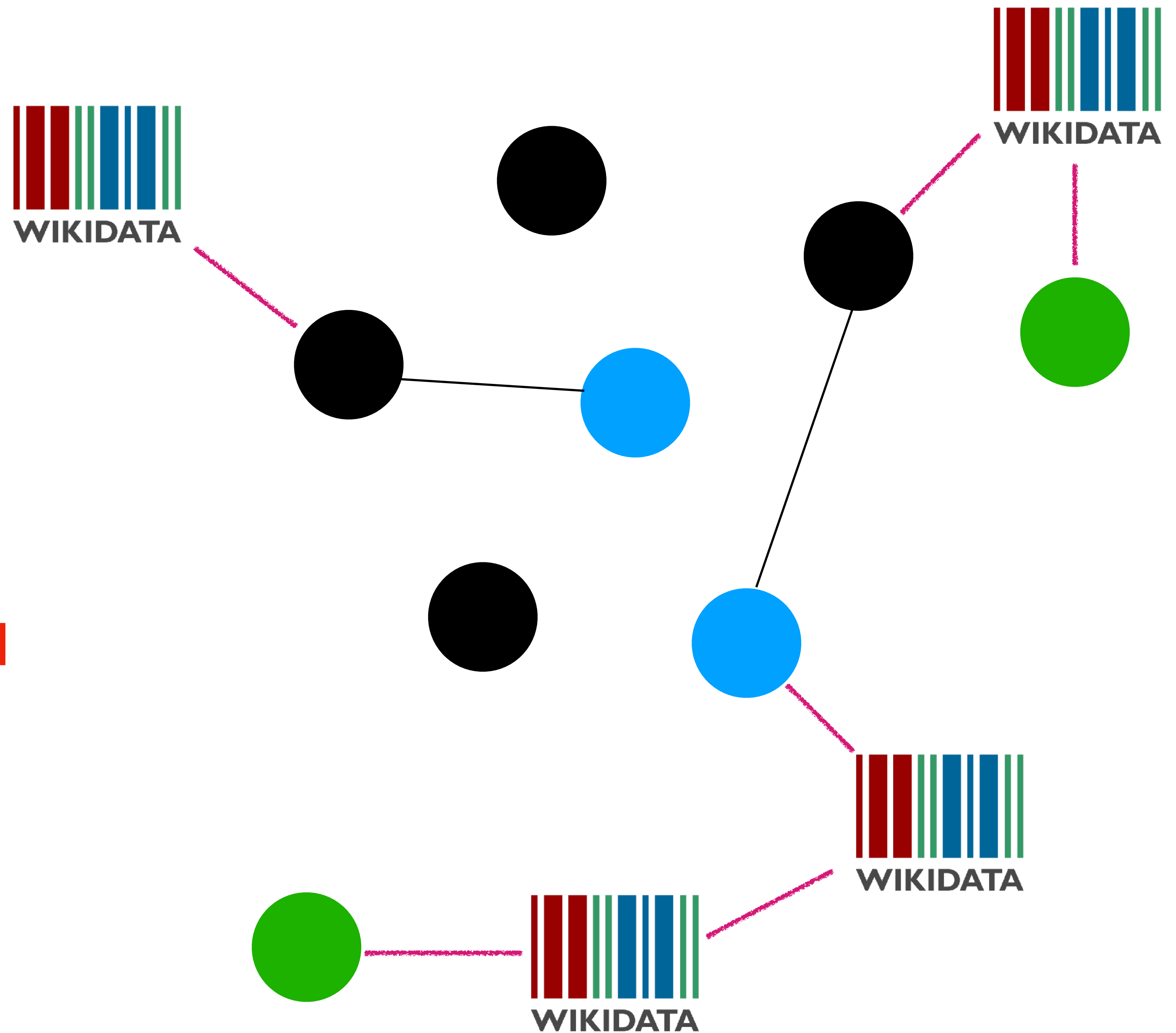
**Small islands of thin data**





# THIS IS OUR COLLECTION CONNECTED TO WIKIDATA..

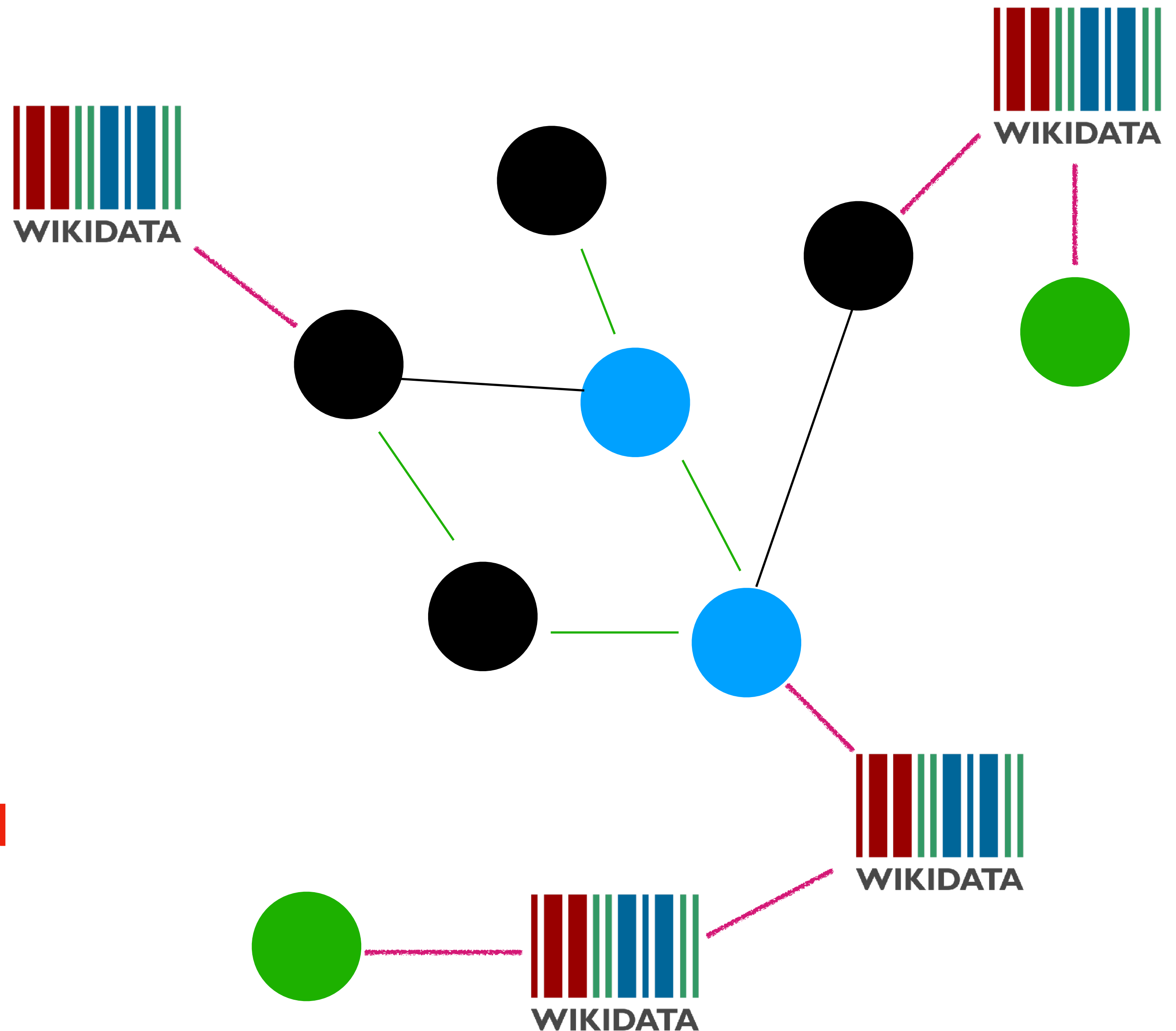
Small islands of **connected**  
data





# THIS IS OUR COLLECTION INTERLINKED VIA INFORMATION EXTRACTION TECHNIQUES..

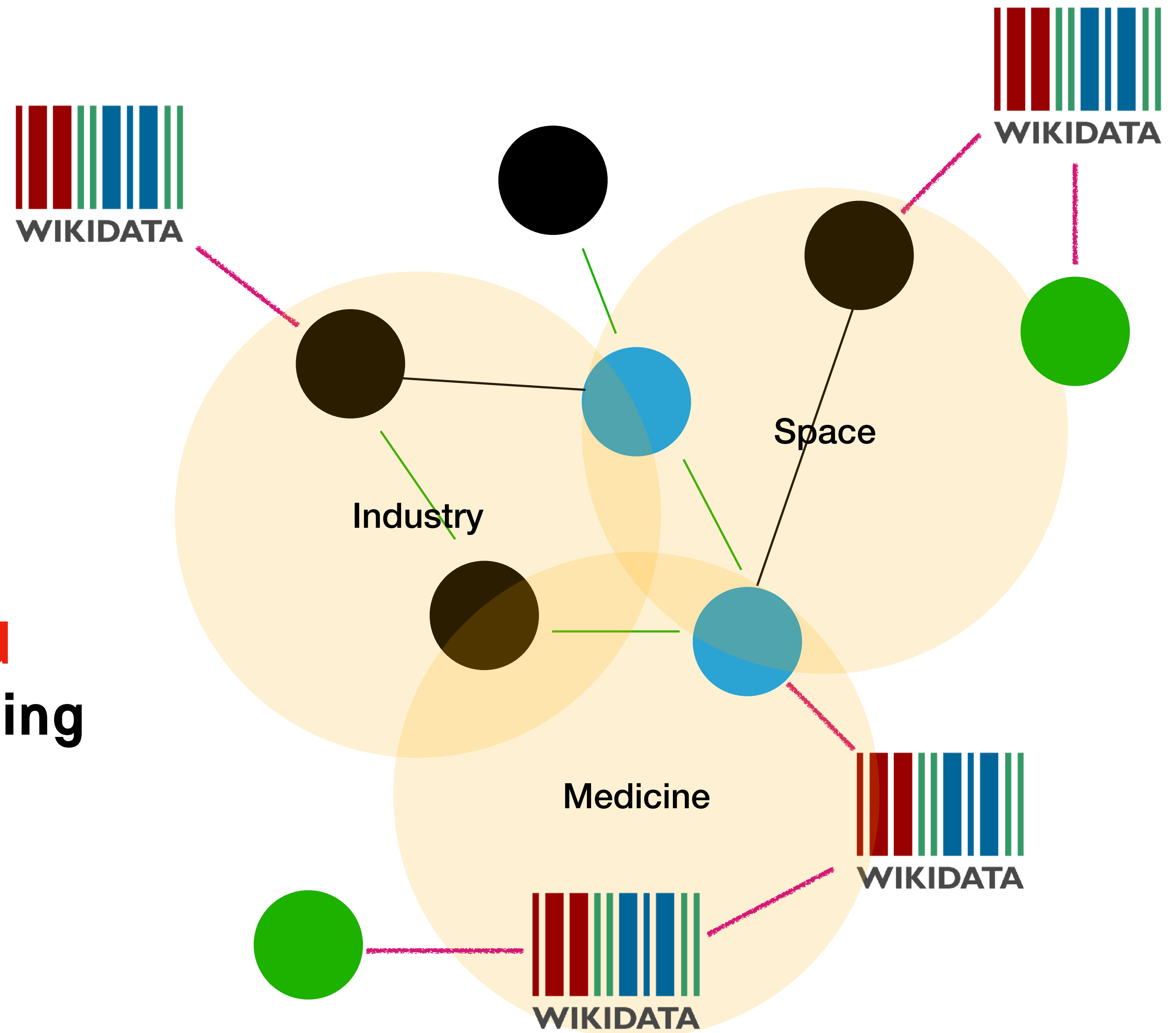
Small islands of **connected**  
and **interlinked** data





# THIS IS OUR COLLECTION WITH NEW GROUPINGS..

Small islands of **connected**  
and **interlinked** data exposing  
new **groupings**

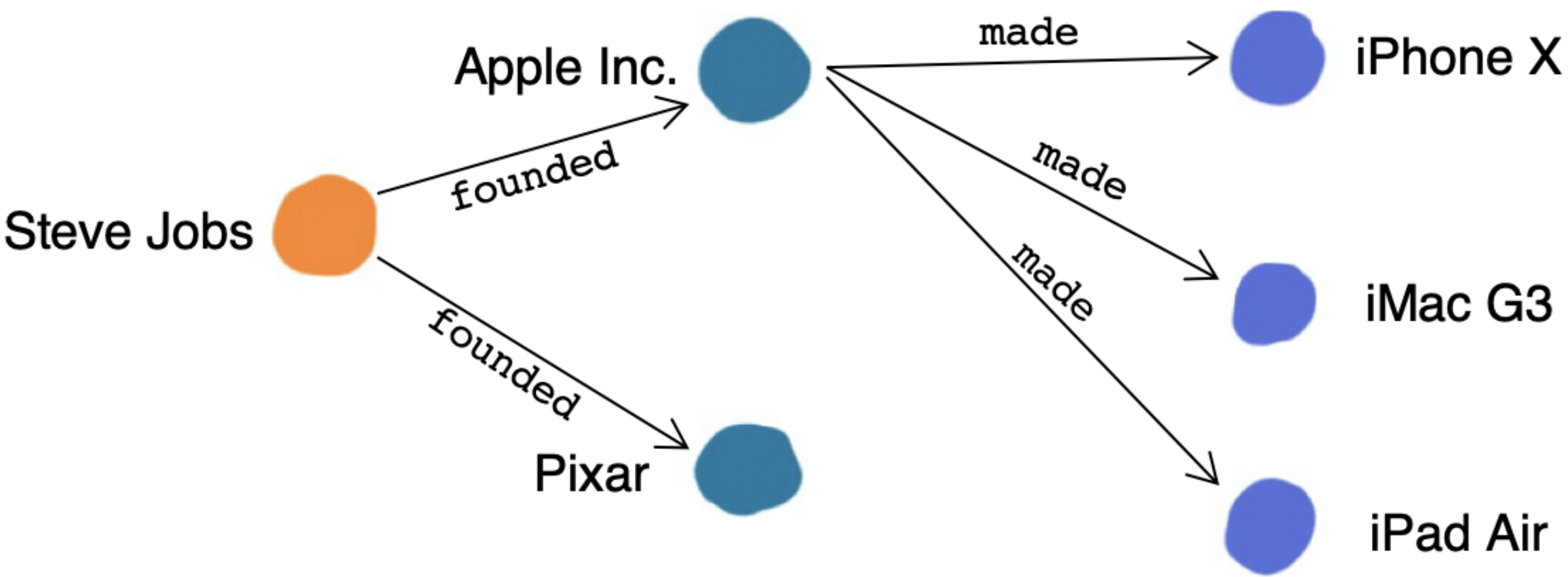
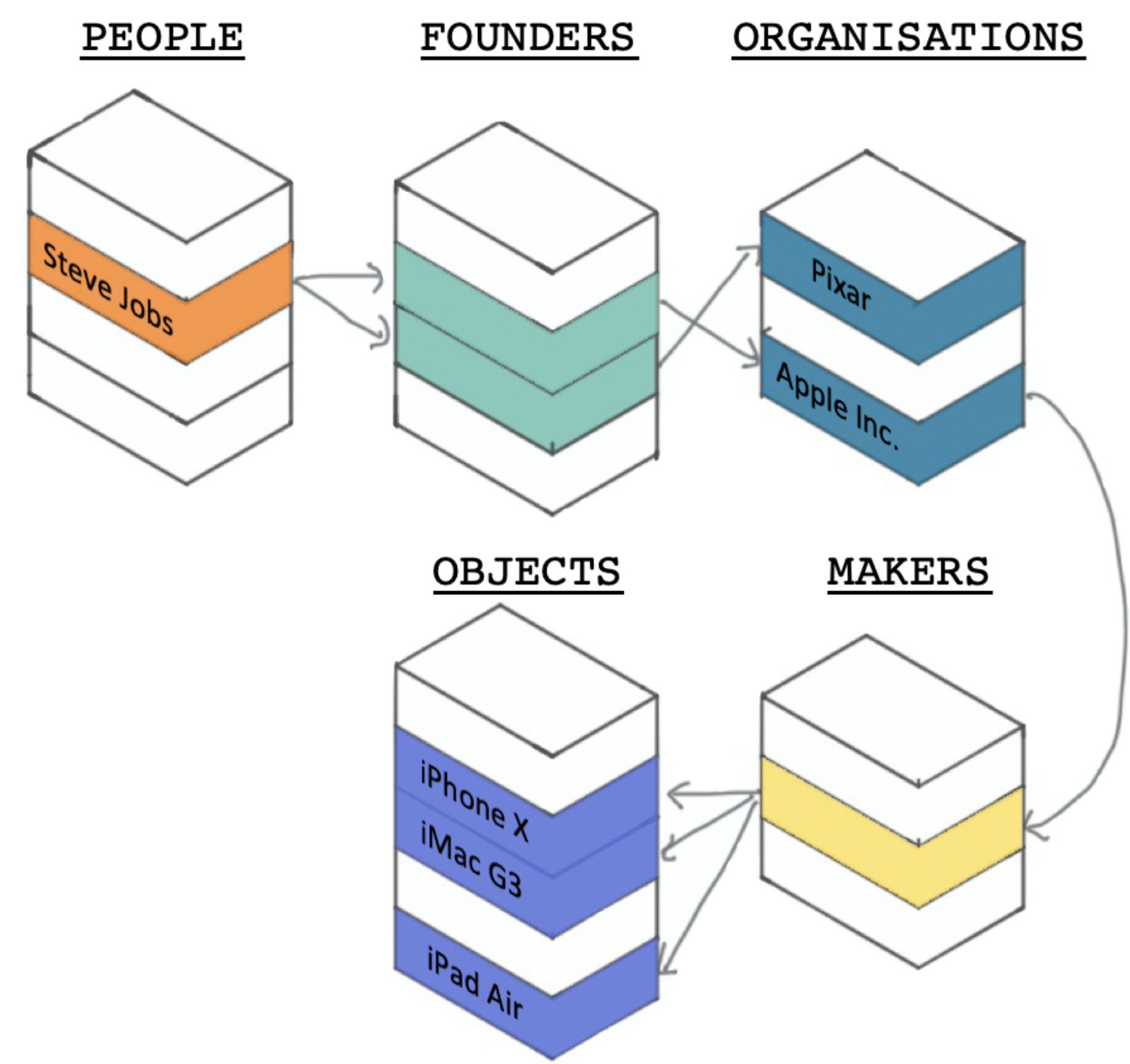


# THREE TECHNOLOGIES

1. Machine Learning (specifically Natural Language Processing)
2. Linked data
3. Knowledge graphs



# LINKED DATA AND KNOWLEDGE GRAPHS

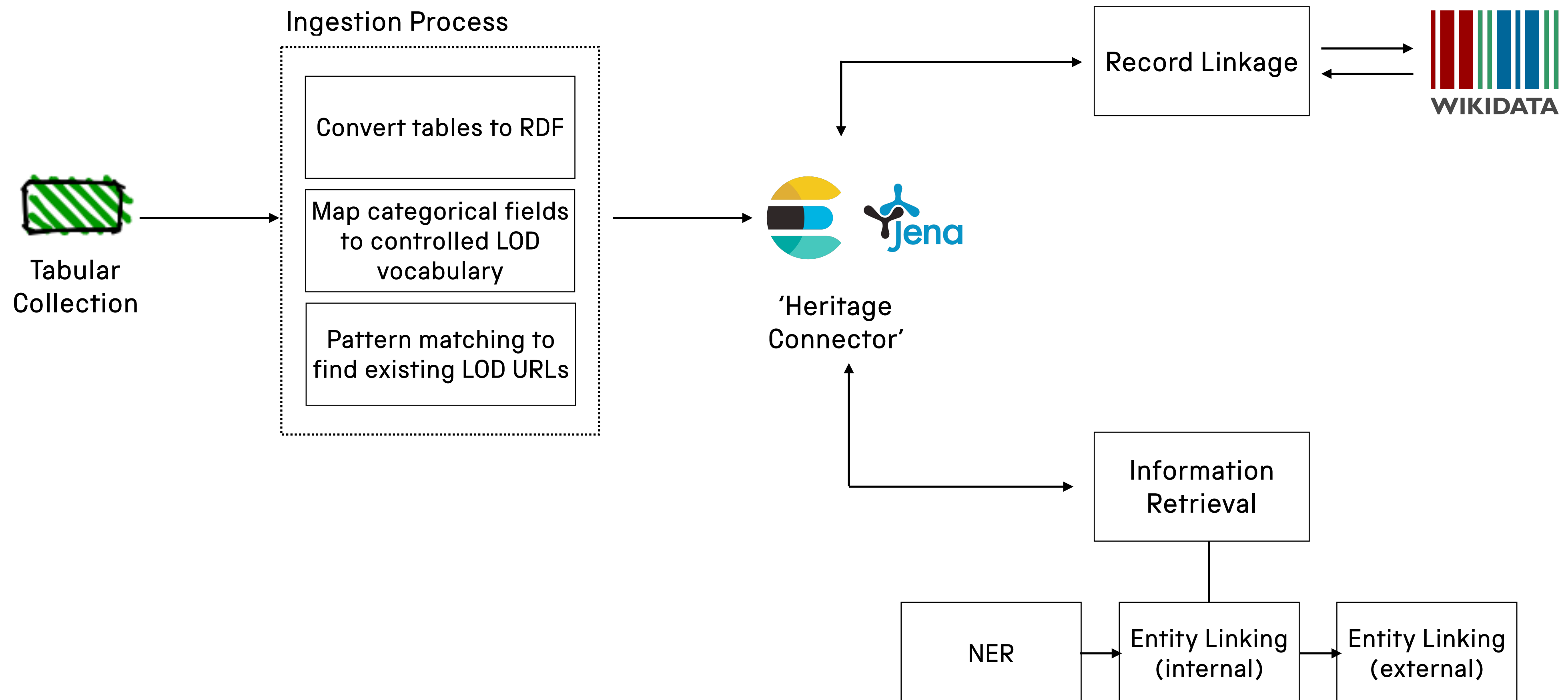


text     Anna Atkins was born in Tonbridge, Kent

triples     anna\_atkins, birth\_place, tonbridge\_kent

linked data     <https://collection.sciencemuseumgroup.org.uk/people/cp11399>:  
<https://www.wikidata.org/wiki/Property:P19>,  
<https://www.wikidata.org/wiki/Q936183>

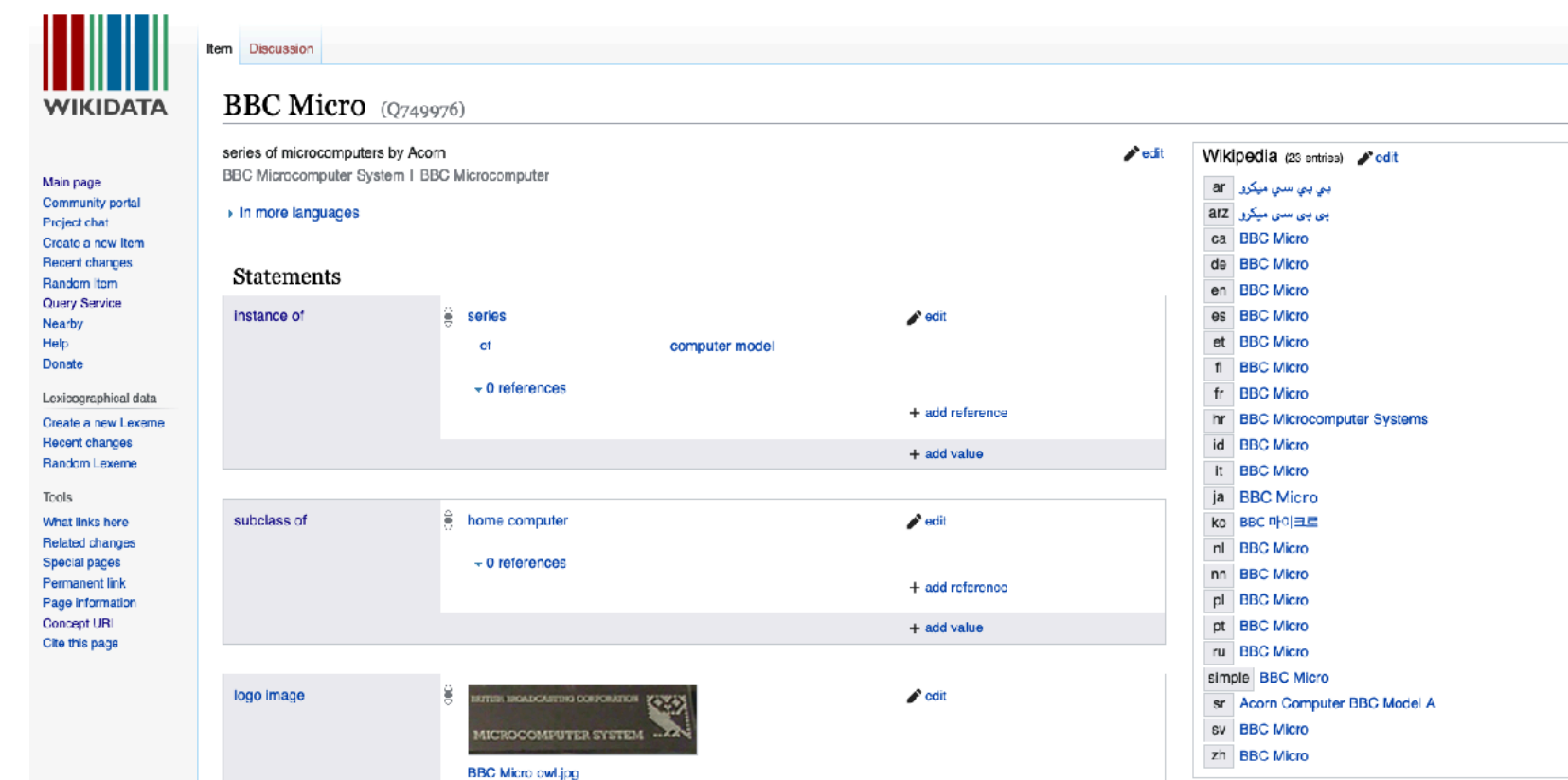
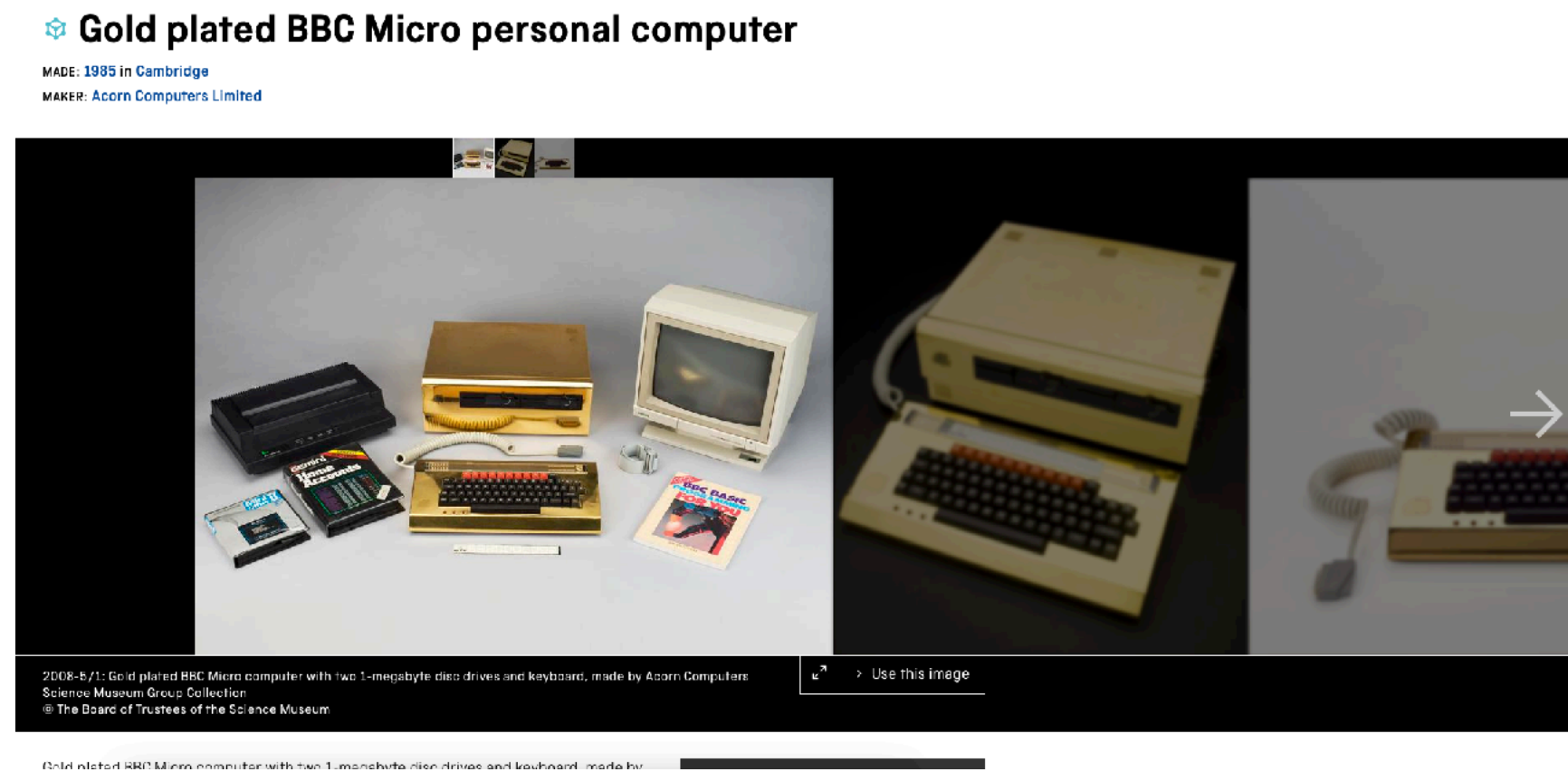
# BUILDING THE KNOWLEDGE GRAPH: TECHNICAL SUMMARY



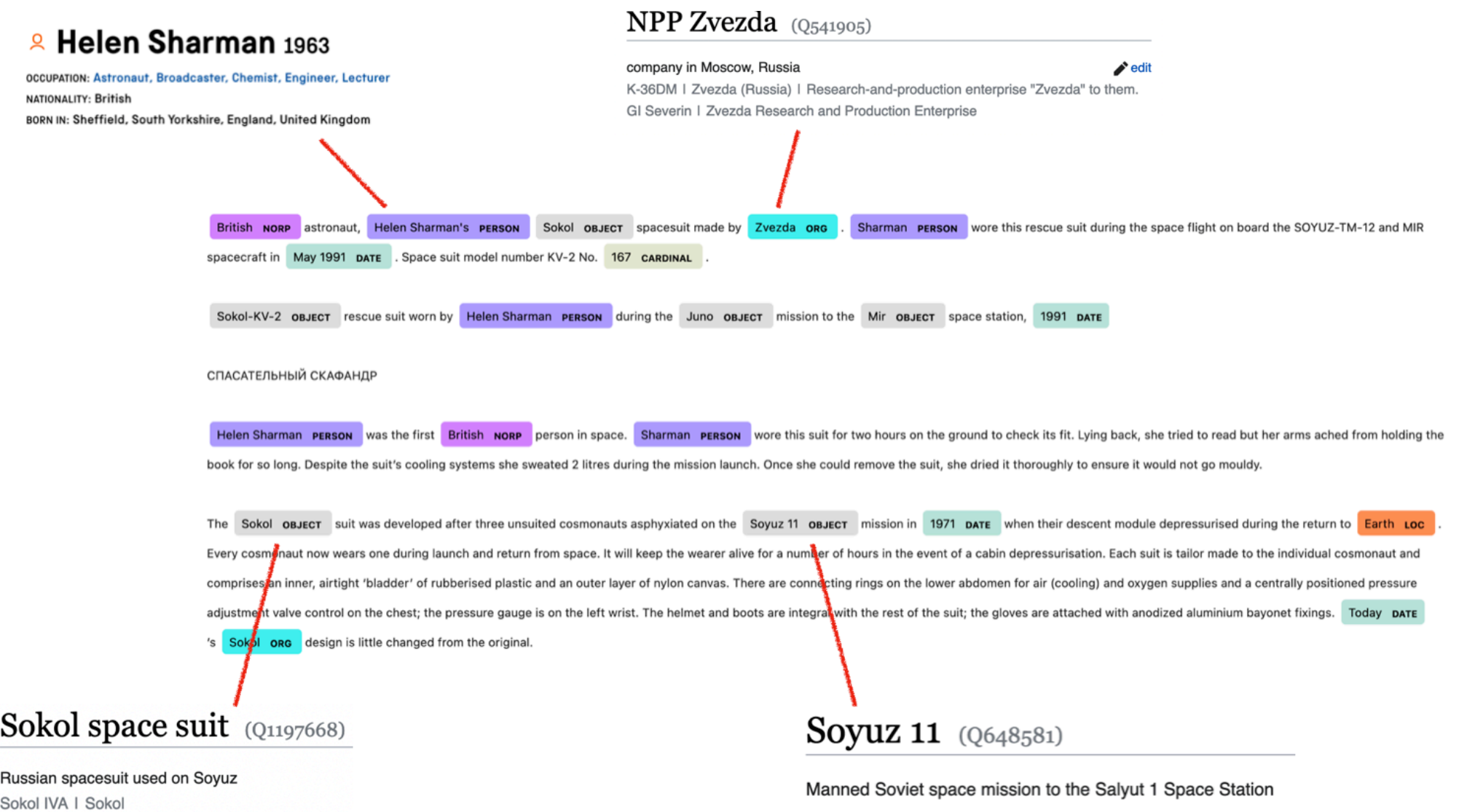


# NATURAL LANGUAGE PROCESSING

## 1. Record Linkage with Wikidata (-> Wikipedia, other collections and databases)



## 2. Named Entity Recognition (NER) and Entity Linking (EL)



# SOME HIGH-LEVEL STATISTICS

- ~354,000 Science Museum Group collection, blog and journal pages and ~325,000 V&A collection records have been processed.
- ~8,700 object, person and organisation have direct connections to Wikidata (only Science Museum Group processed).
- Using NER and EL:
  - ~484,000 connections have been added between collection records,
  - ~648,000 connections have been added to Wikidata,
  - ~1.2 million connections have been added to yet unknown entities.
- In total, there are ~110,000 Wikidata/Wikipedia pages in the knowledge graph.



# USER INTERFACES

# AFFORDANCES OF THIS NEW DATASET

- **macro-views** of the entire collections (plus the blog and journal).
- **serendipitous discovery**, through linking of previously isolated records.
- **new entry points** into the collection from Wikipedia and Wikidata (events, bands, etc.).
- **new forms of interface**, to provide an alternative to keyword search.



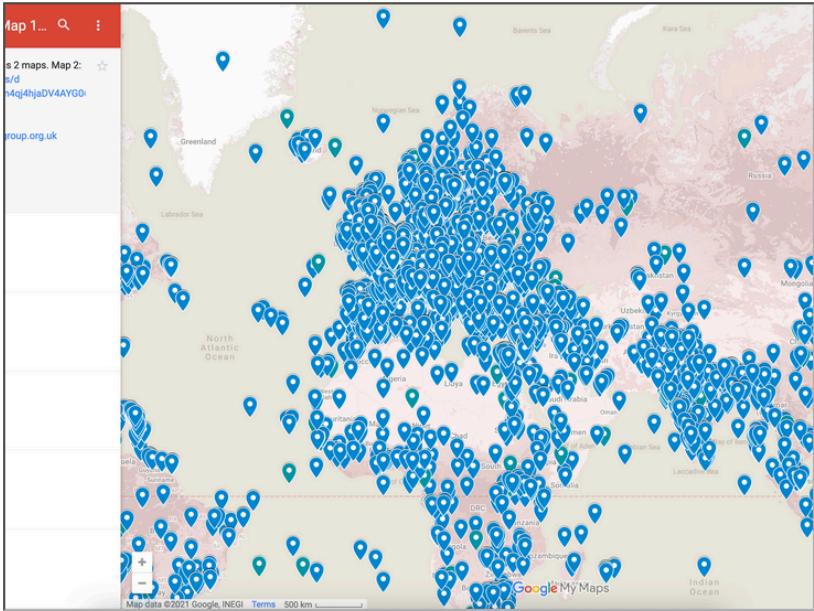
# DEMONSTRATORS

Heritage Connector Demonstrators

heritageconnector.sciencemuseum.org.uk

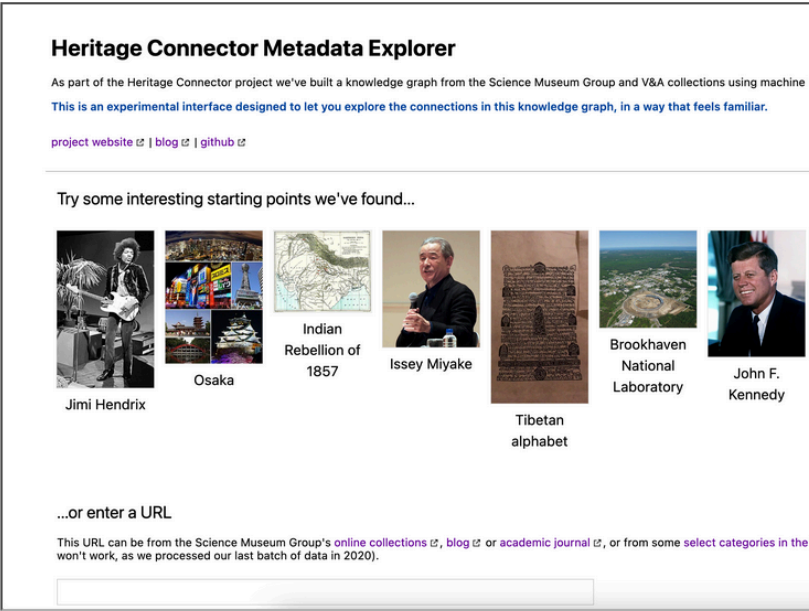
## Heritage Connector Demonstrators

This page serves as a directory of demos for the [Heritage Connector project](#). The source code for most of these demos can be found on [Github](#).



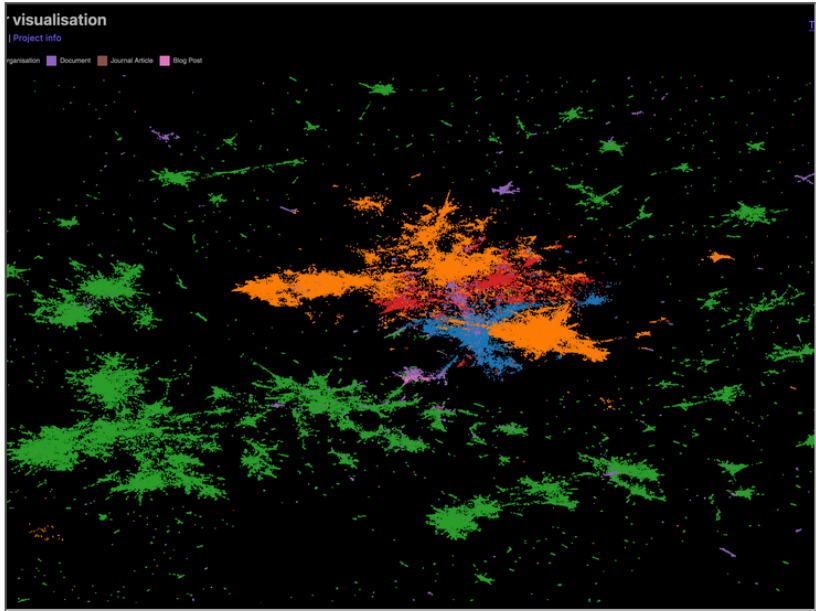
### Mapping locations

Creating Google maps from places found in the Science Museum Group collection.



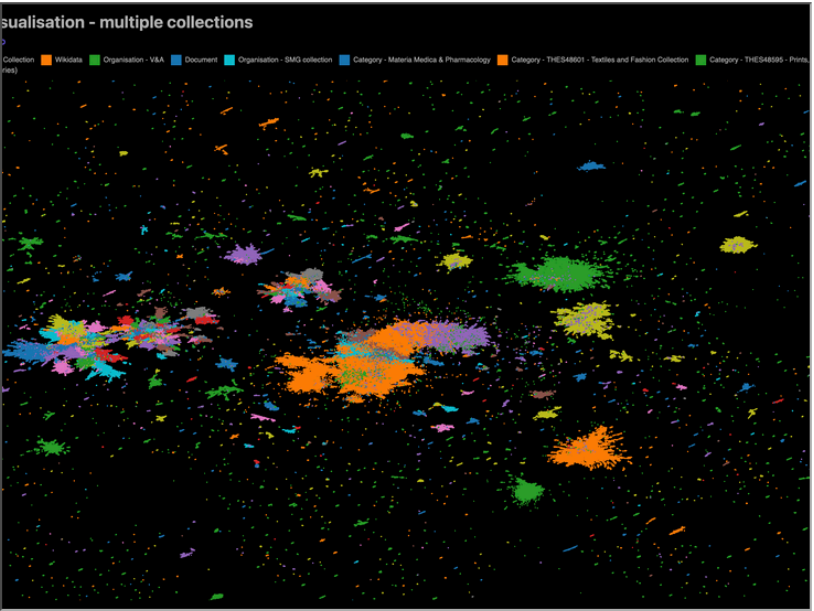
### Metadata explorer

An interface for guided exploration of the Heritage Connector knowledge graph.



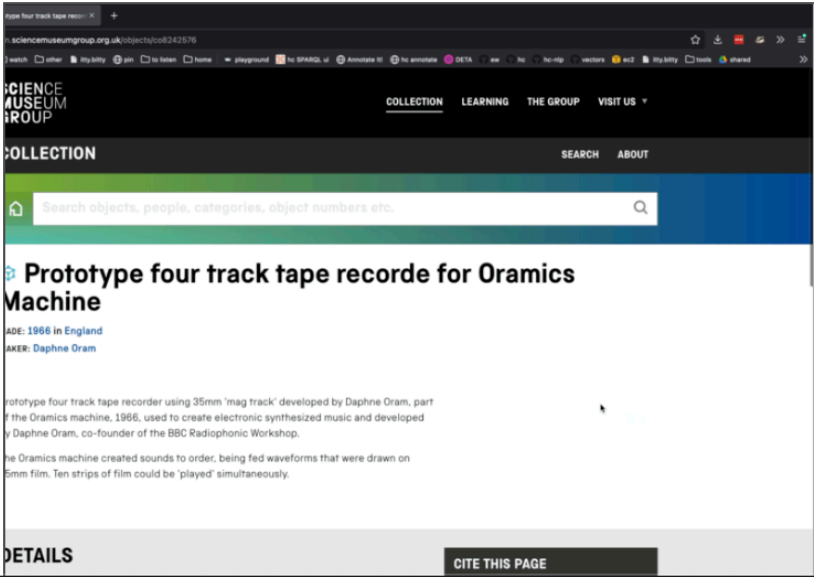
### Visualisation

View a map of a knowledge graph containing the Science Museum Group's collection, blog and journal.

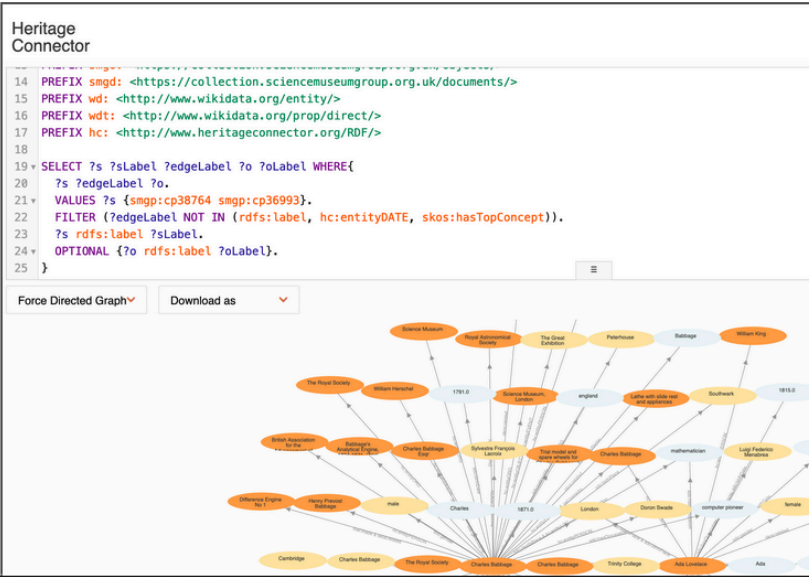


### Visualisation - multiple collections

View a map of a knowledge graph containing the Science Museum Group's collection, blog and journal, and part of the V&A's collection.



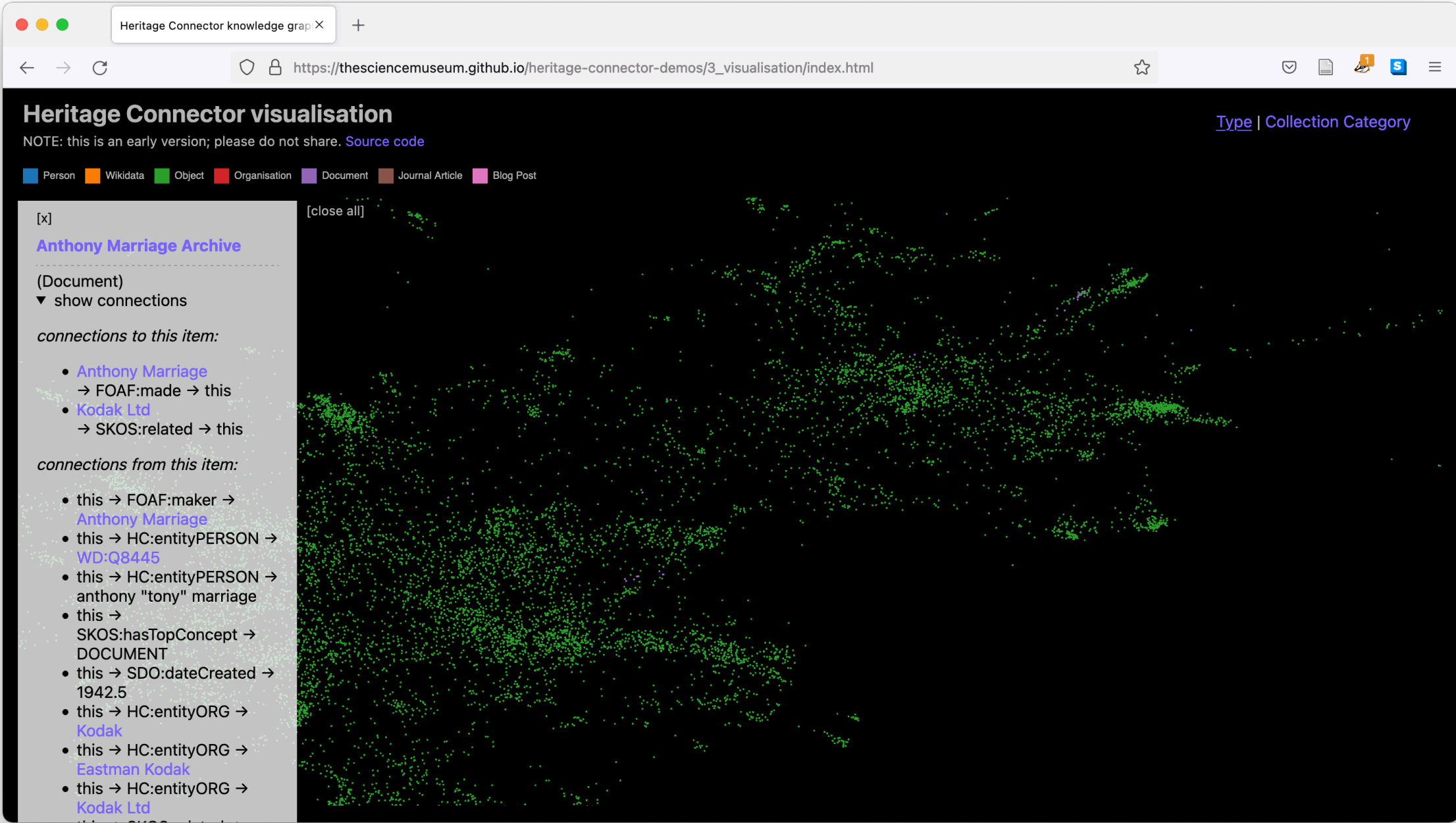
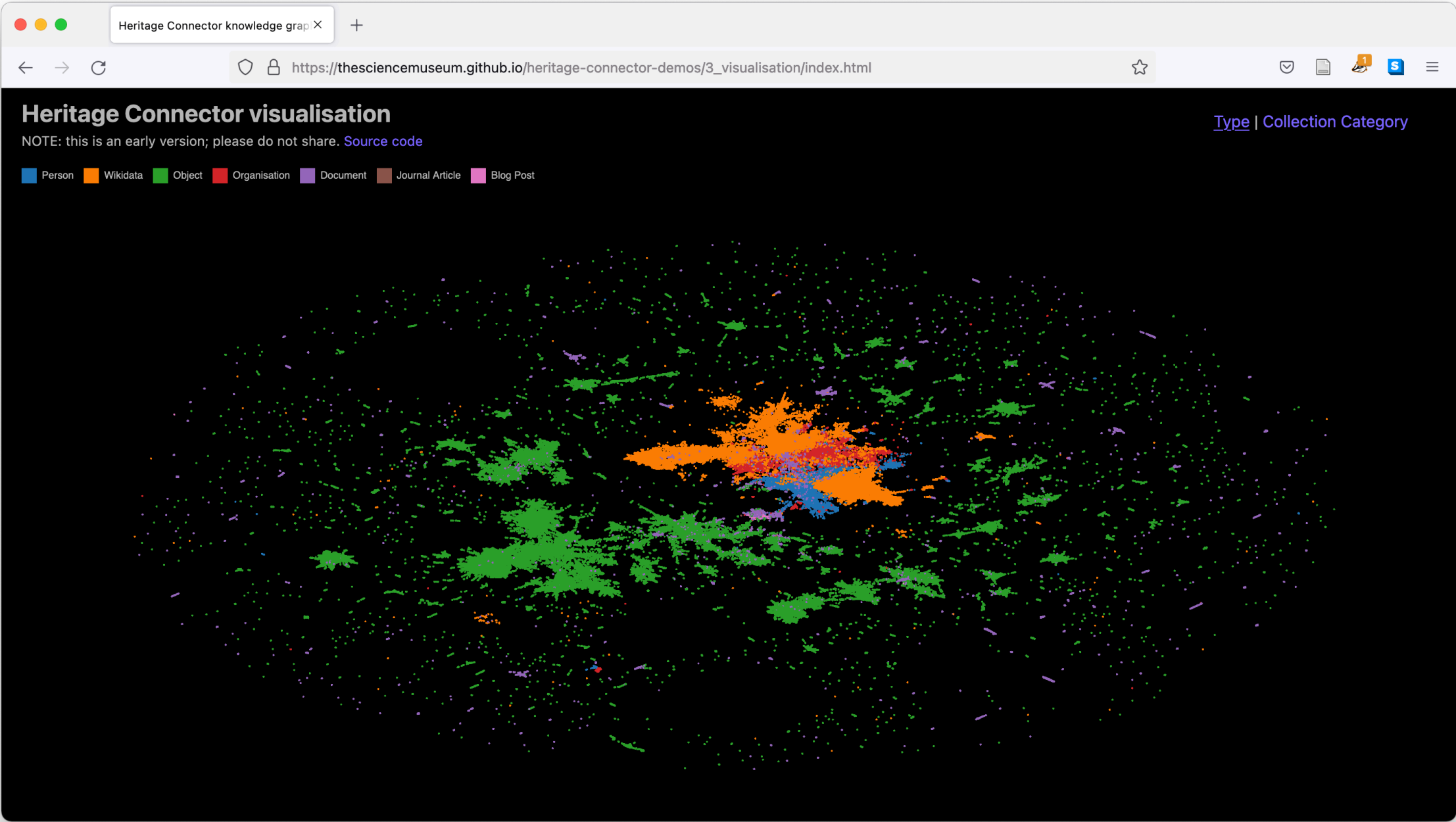
### Collection



### Heritage Connector



# MACRO COLLECTION VISUALISATION





# BOOKMARKLET


SCIENCE  
MUSEUM  
GROUP

COLLECTION


LEARNING

THE

COLLECTION



Search objects, people, categories, object numbers etc.




Spirit of the Southern

MADE: 1969, about in [England](#)

COMMISSIONER: [British Railways, Southern Region](#)

ARTIST: [Harry Stevens](#)



Heritage Connector - connections found

Close [x]

[Log interesting/incorrect connections \(HC team\)](#)

EXISTING RECORD METADATA

RDFS:label

- Spirit of the Southern

HC:database

- mimsy

RDF:type

- [WD:Q3305213](#)

SKOS:hasTopConcept

- OBJECT

SDO:isPartOf

- Pictorial Collection (Railway)

XSD:additionalType

- painting

SDO:material

- acrylic paint

SDO:dateCreated

- 1969.0

EXISTING LINKS TO OTHER PAGES

FOAF:maker

- [British Railways, Southern Region \[SMGP:cp44102\]](#)
- [Harry Stevens \[SMGP:cp38391\]](#)

SAMEAS/SIMILAR LINKS

FROM NER & NEL

HC:entityPERSON

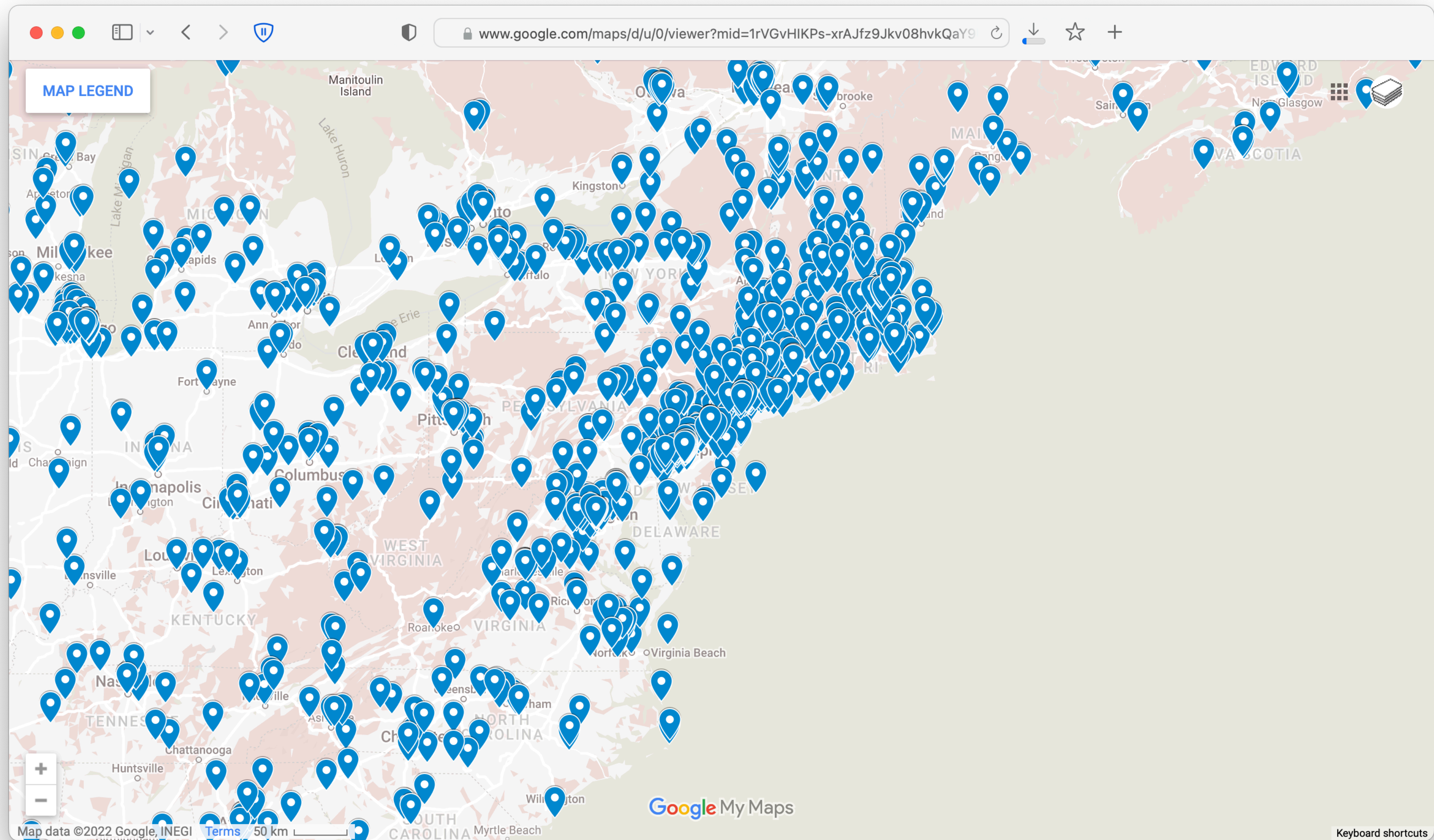
- [I.w ibbotson](#)
- [Harry Stevens \[SMGP:cp38391\]](#)

HC:entityORG

- [British Railways \[SMGP:cp3862\]](#)
- [British Railways \[SMGP:ap315\]](#)
- [Southern Region of British Railways \[WD:Q7570287\]](#)



# MAP





# FINDINGS...

- Methods used can generate huge numbers of links.
- Fruitful avenues for improved search and discovery.
- Collection text fields and article content are valuable locations for link building.
- Opportunities in new data alongside collection catalogue.
- False positive results are usually readily apparent.
- New forms of interface can be developed (and are needed).

# THINGS TO CONSIDER...

- Machine learning generated content needs framing/ contextualisation.
- Challenges cultural heritage notions of “canonical” collection catalogue data.
- Important to approach critically.
- Specialist technical skills needed.
- Human expert input needed throughout.



# THANKS

<https://www.sciencemuseumgroup.org.uk/project/heritage-connector/>

<https://thesciemuseum.github.io/heritageconnector/>

**SCIENCE  
MUSEUM  
GROUP**

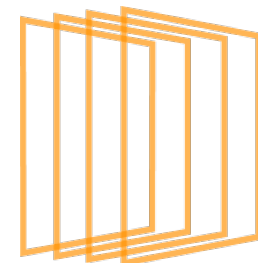


**SCHOOL OF  
ADVANCED STUDY  
UNIVERSITY  
OF LONDON**

**V&A**



**Arts and  
Humanities  
Research Council**



**TOWARDS  
A NATIONAL  
COLLECTION**

**cogapp**



**WIKIDATA**



**WIKIMEDIA  
UK**