Heritage Connector

Jamie Unwin & Kalyan Dutia

SCIENCE MUSEUM GROUP

UKRI Arts and Humanities Research Council

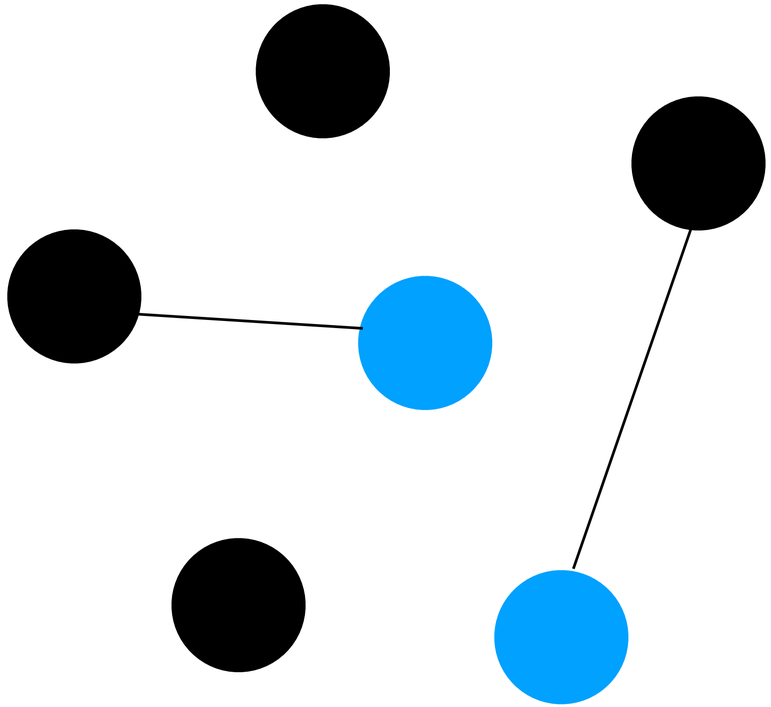
SCHOOL OF ADVANCED STUDY UNIVERSITY OF LONDON

V&A

“TRANSFORMING TEXT INTO DATA TO EXTRACT MEANING AND MAKE CONNECTIONS”

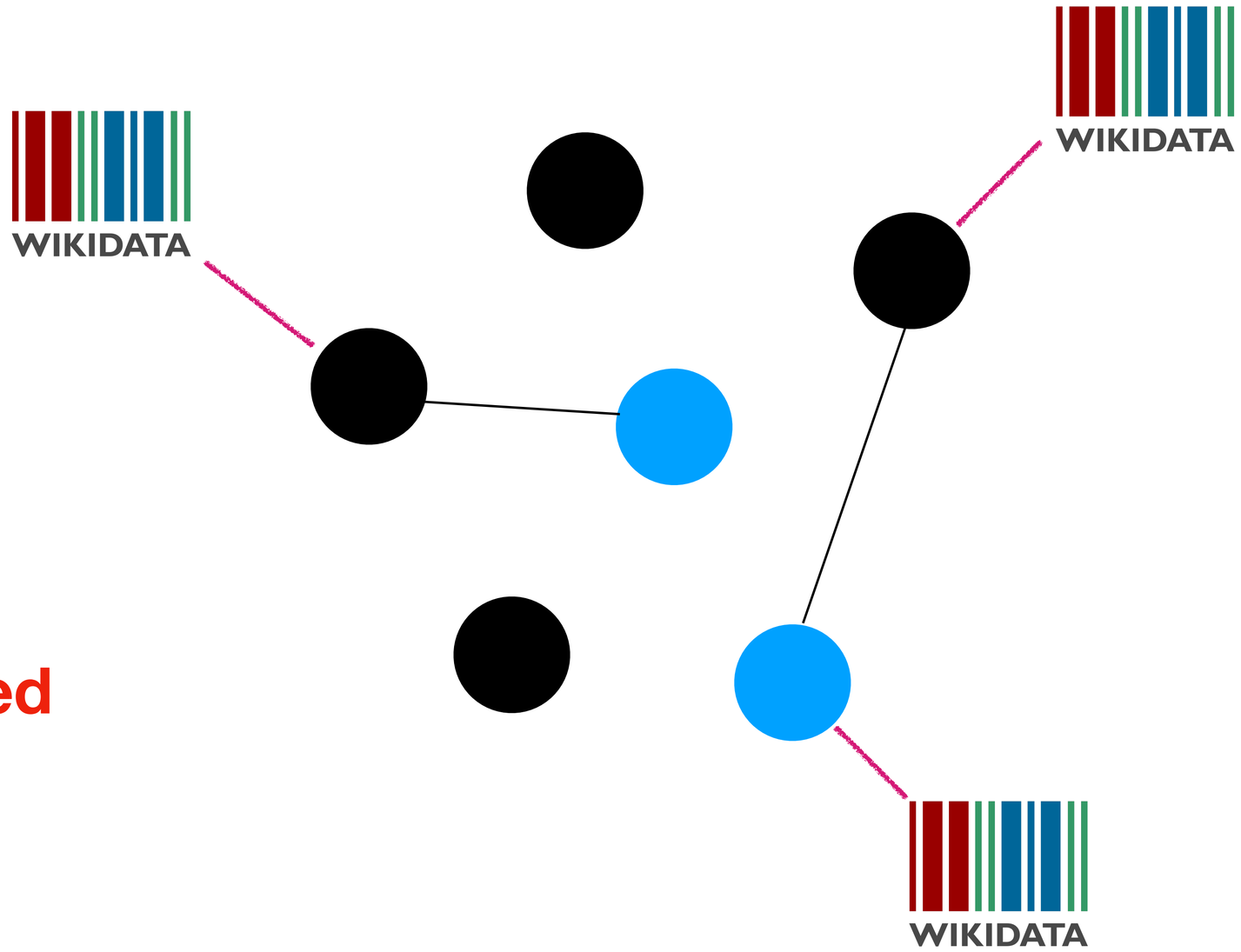
This is our collection now..

Small islands of thin data



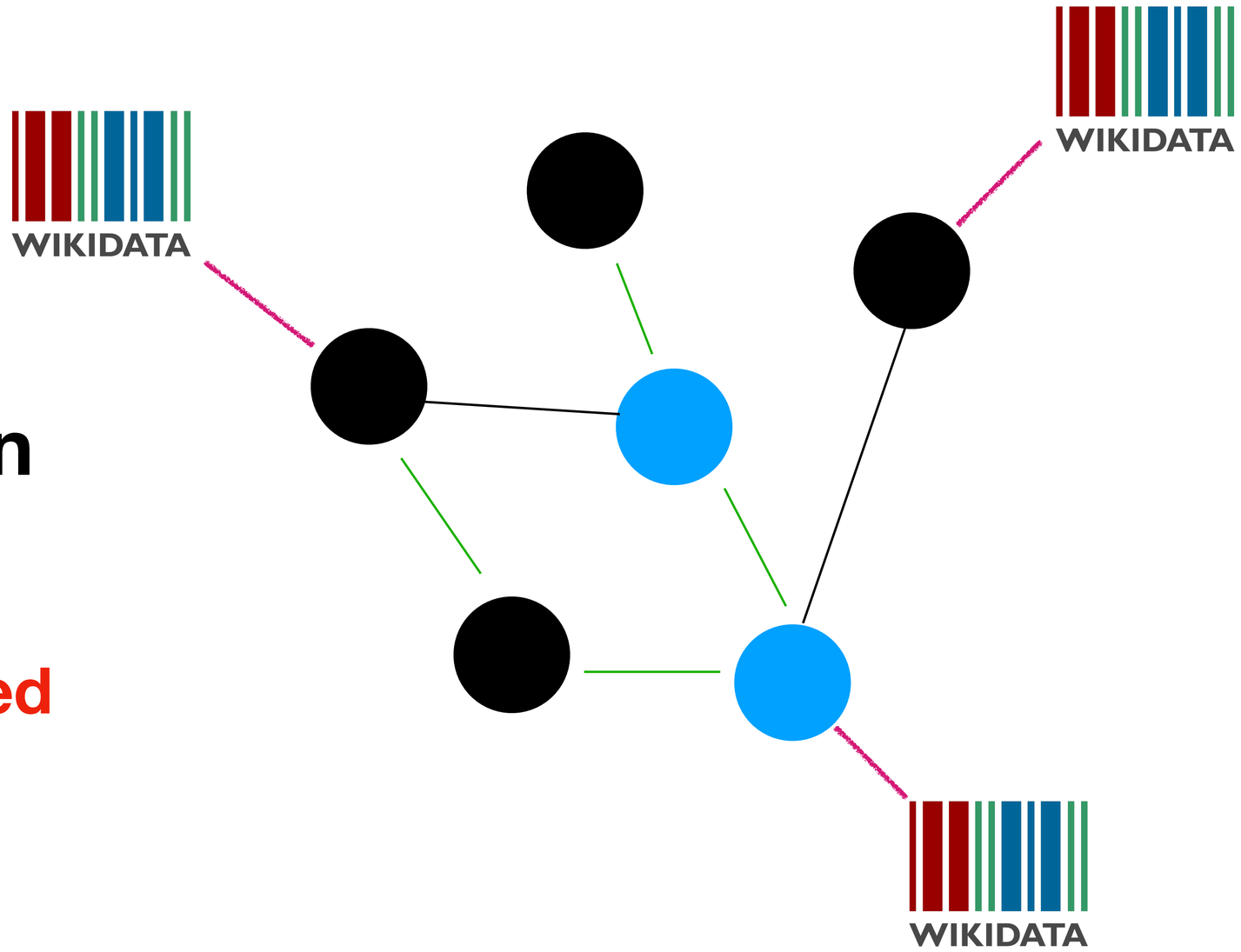
This is our collection connected to Wikidata..

Small islands of connected data



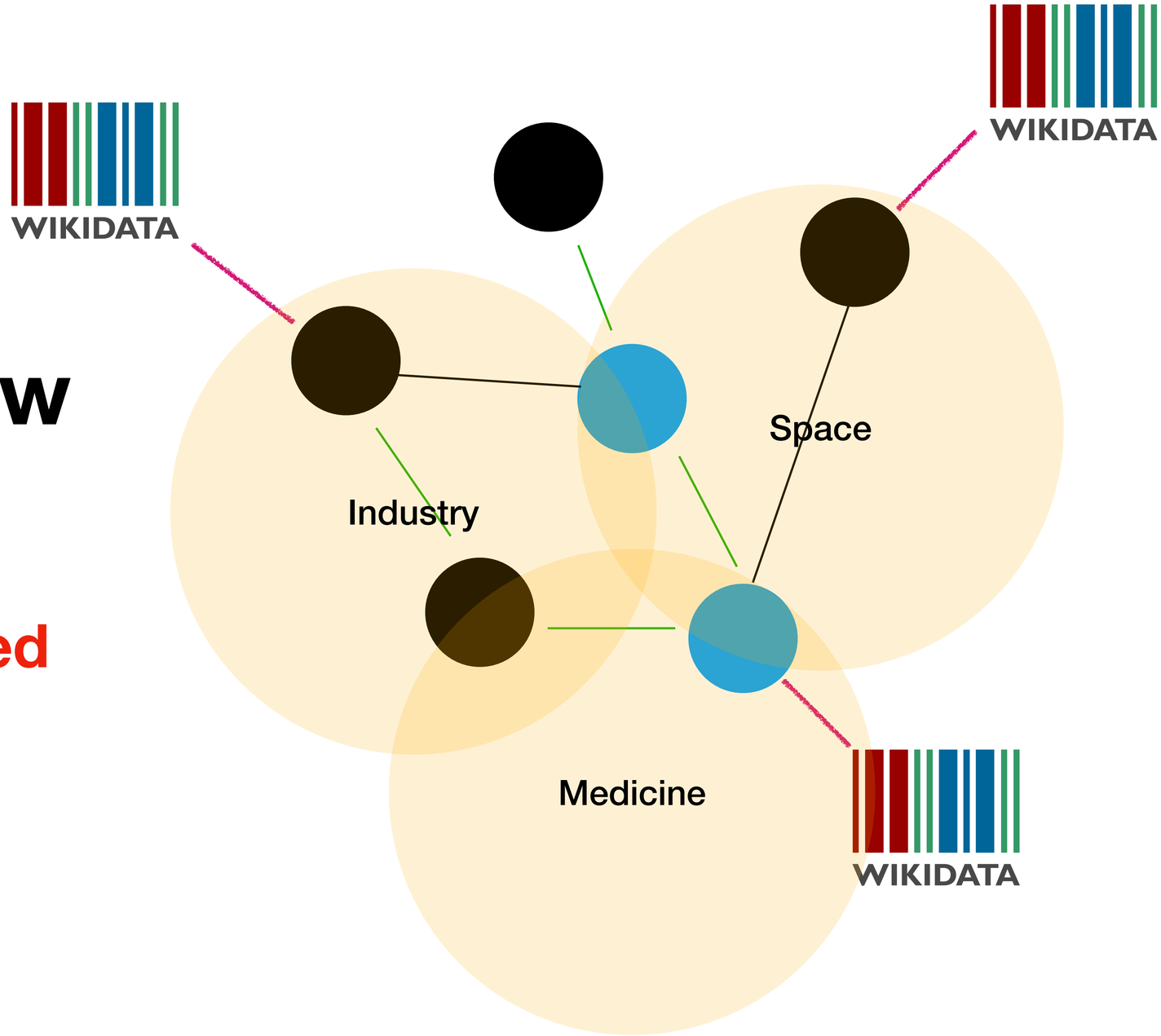
This is our collection interlinked via information extraction techniques..

Small islands of connected and interlinked data



This is our collection with new groupings..

Small islands of connected and interlinked data exposing new groupings



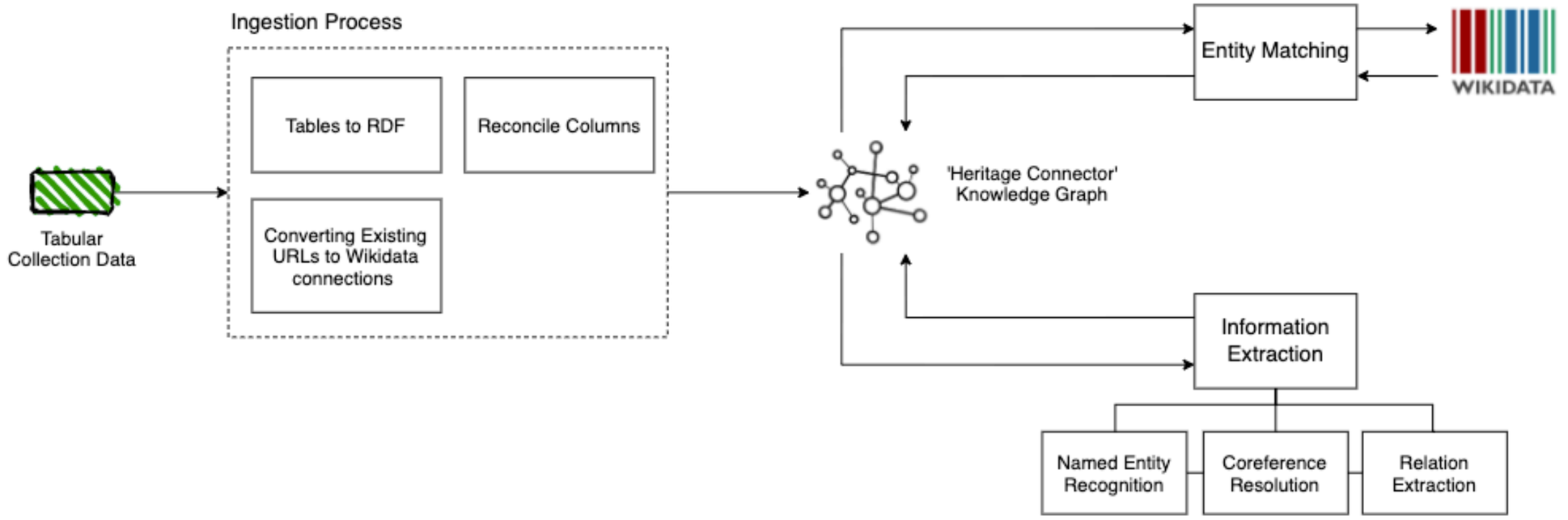
Why are we doing this?

- Human resources are limited, especially expert ones
- Introduce new topics and themes not inherent in our current data
- Enrich specifics types of records/objects with additional data
- Stop our record pages becoming ‘dead ends’ to users.
- Along with all the other obvious benefits of LoD :-)

Things were thinking about

- What techniques are be best used to build these new relationships and groupings at scale?
- How might confidence in these relationships impact on their usefulness?
- Where is the best use of human input in supporting such an approach?
- What gaps and biases emerge when these relationships are created, and which hitherto unexpected connections are made?

How are we doing this?


Ingestion Process
Tables to RDF
Reconcile Columns
Converting Existing URLs to Wikidata connections
Tabular Collection Data
'Heritage Connector' Knowledge Graph
Entity Matching
WIKIDATA
Information Extraction
Named Entity Recognition
Coreference Resolution
Relation Extraction

Information Extraction within Heritage Connector

Named Entity Recognition (NER)

Coreference Resolution

Relation Extraction

“British astronaut, Helen Sharman’s Sokol spacesuit made by Zvezda. Sharman wore this rescue suit during the space flight on board the SOYUZ-TM-12 and MIR spacecraft in May 1991.”

Information Extraction within Heritage Connector

Named Entity Recognition (NER)

Coreference Resolution

Relation Extraction

“British NORP astronaut, Helen Sharman’s PERSON Sokol spacesuit OBJECT made by Zvezda ORG. Sharman PERSON wore this rescue suit during the space flight on board the SOYUZ-TM-12 OBJECT and MIR spacecraft OBJECT in May 1991 DATE.”

Information Extraction within Heritage Connector

Named Entity Recognition (NER)

Coreference Resolution

Relation Extraction

“British astronaut, Helen Sharman’s PERSON Sokol spacesuit made by Zvezda. Sharman PERSON wore this rescue suit during the space flight on board the SOYUZ-TM-12 and MIR spacecraft in May 1991 DATE.”

```
SM1011: (‘Helen Sharman’, ‘Sharman’)
```

Information Extraction within Heritage Connector

Named Entity Recognition (NER)

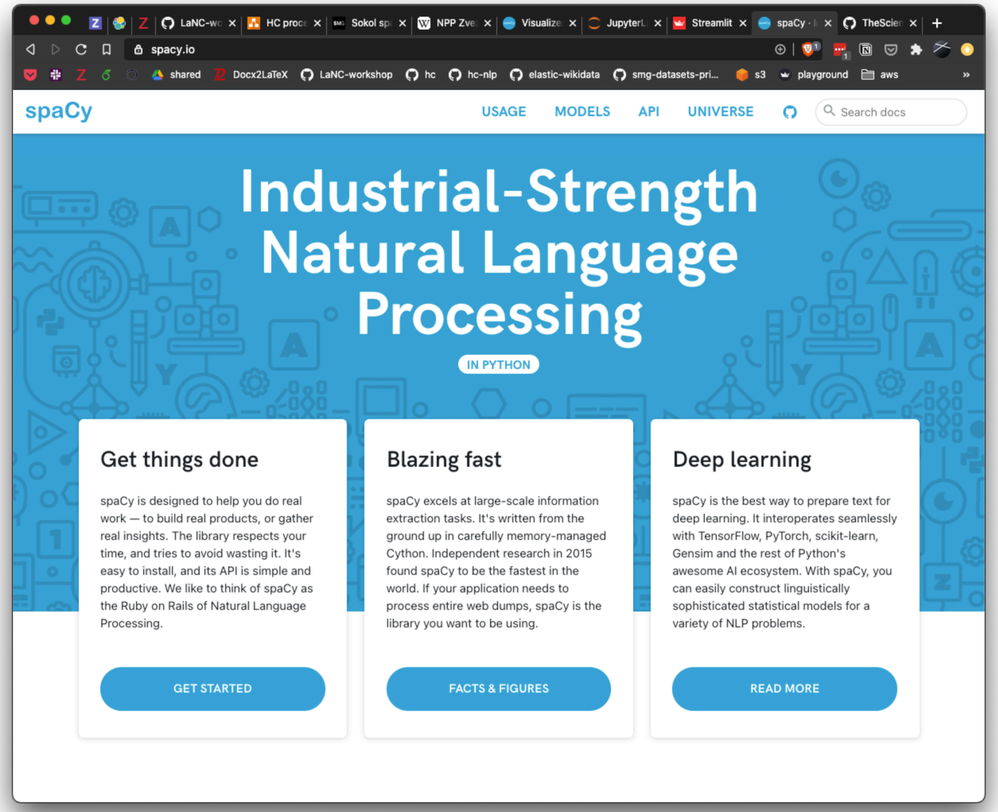
Coreference Resolution

Relation Extraction

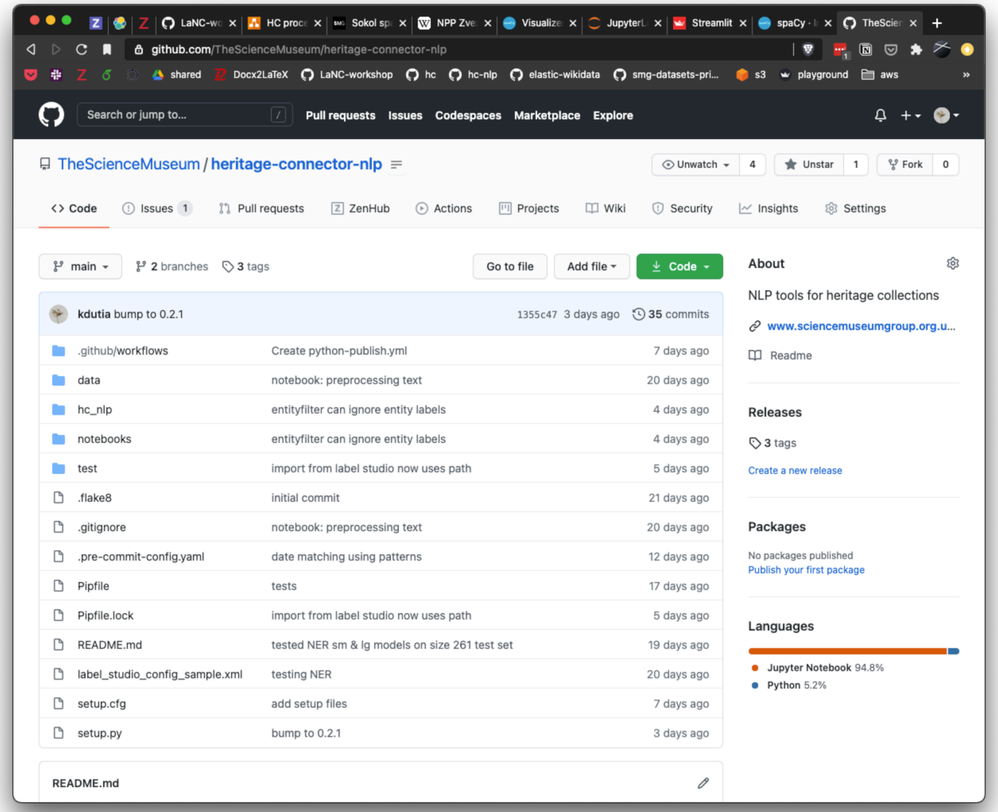
“British astronaut, Helen Sharman’s Sokol spacesuit OBJECT made by REL Zvezda ORG. Sharman wore this rescue suit during the space flight on board the SOYUZ-TM-12 and MIR spacecraft in May 1991.”

```
('Sokol spacesuit', made_by, 'Zvezda')
      SM1013                  SM1012
```

Demo: Test & Play



spacy.io



github.com/TheScienceMuseum/heritage-connector-nlp

spaCy Default Pipeline

```
nlp = spacy.load("en_core_web_md")
doc = nlp(text)
```

text

annotated

“British astronaut, Helen Sharman’s Sokol spacesuit made by Zvezda. Sharman wore this rescue suit during the space flight on board the SOYUZ-TM-12 and MIR spacecraft in May 1991.”

“British astronaut, Helen Sharman’s Sokol spacesuit made by Zvezda. Sharman wore this rescue suit during the space flight on board the SOYUZ-TM-12 and MIR spacecraft in May 1991.”

Heritage Connector Components

For efficiently applying expert knowledge to the entity recognition process

Thesaurus Matcher

Finds occurrences of entities in text based on **lookup in an external thesaurus (or gazetteer)**

Rule-based Matcher

Finds occurrences of entities based on **syntactic patterns in text**

Using a Gazetteer with

Thesaurus Matcher

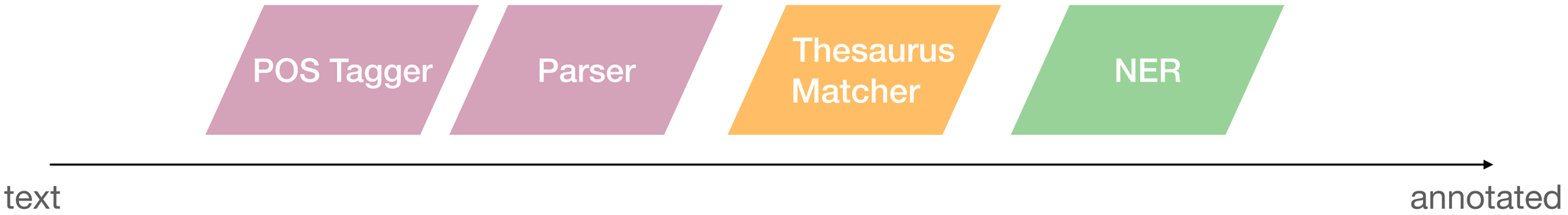
```
{"label": "GPE", "pattern": "Laceby"}
{"label": "GPE", "pattern": "Denby"}
{"label": "GPE", "pattern": "Hauxwell"}
{"label": "GPE", "pattern": "Rudgwick"}
{"label": "GPE", "pattern": "Oving"}
{"label": "GPE", "pattern": "Hatley St. George"}
{"label": "GPE", "pattern": "Muchland"}
{"label": "GPE", "pattern": "Woodford near Thrapstone"}
{"label": "GPE", "pattern": "Coleby"}
{"label": "GPE", "pattern": "Cwm B\ufffdch"}
{"label": "GPE", "pattern": "Efenechtyd"}
{"label": "GPE", "pattern": "Wenden"}
{"label": "GPE", "pattern": "Donington"}
```

gazetteer.jsonl
16,032 terms

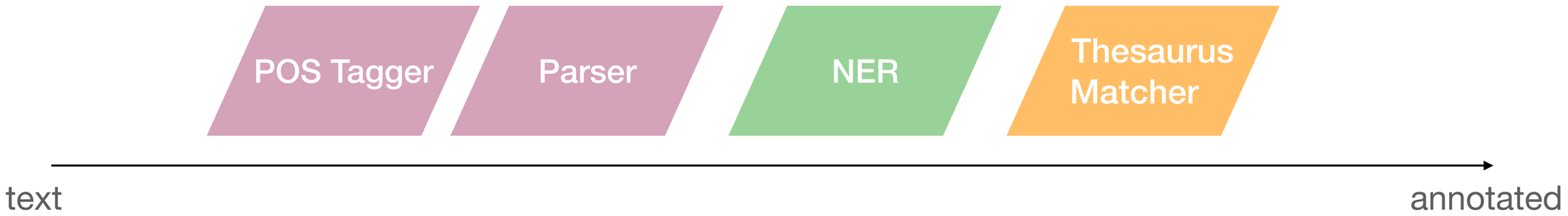
```
# create ThesaurusMatcher
thesaurus_matcher = ThesaurusMatcher(nlp,
                                     thesaurus_path='./gazetteer.jsonl',
                                     case_sensitive=False)

# add to spaCy pipeline
nlp.add_pipe(thesaurus_matcher, before='ner')
```

Easy to add to spaCy pipeline

1. Gazetteer before NER



2/3. Gazetteer after NER, with or without overwrite



Demo link:
https://github.com/LinkedPasts/LaNC-workshop/tree/main/heritageconnector