

Encyclopædia Britannica;

James OR, A Fullerton

DICTIONARY

O F

A R T S and S C I E N C E S,

COMPILED UPON A NEW PLAN.

IN WHICH

The different SCIENCES and ARTS are digested into
distinct Treatises or Systems;

A N D

The various TECHNICAL TERMS, &c. are explained as they occur
in the order of the Alphabet.

ILLUSTRATED WITH ONE HUNDRED AND SIXTY COPPER

By a SOCIETY of GENTLEMEN in SCOTLA

IN THREE VOLUMES

VOL. I.

EDINBURGH

Printed for A. BELL and C. M

And sold by COLIN MACFARQUHAR, at

M.DCC

*frances: An AI-toolbox to
discover automatically insights
from Data Foundry collections*

Dr. Rosa Filgueira,
Assistant Professor,
Heriot-Watt University,
Email: R.Filgueira@hw.ac.uk

*frances**: New AI-toolbox to discover automatically insights

New ways to unlock the full value of NLS digital collections

- Objective 1: New facilities to run more complex text analysis queries → ML/NLP techniques.
- Objective 2: Full integration with NLS Data Foundry → Hide the large-scale text mining complexity
- Using the **Encyclopaedia Britannica** as the core dataset

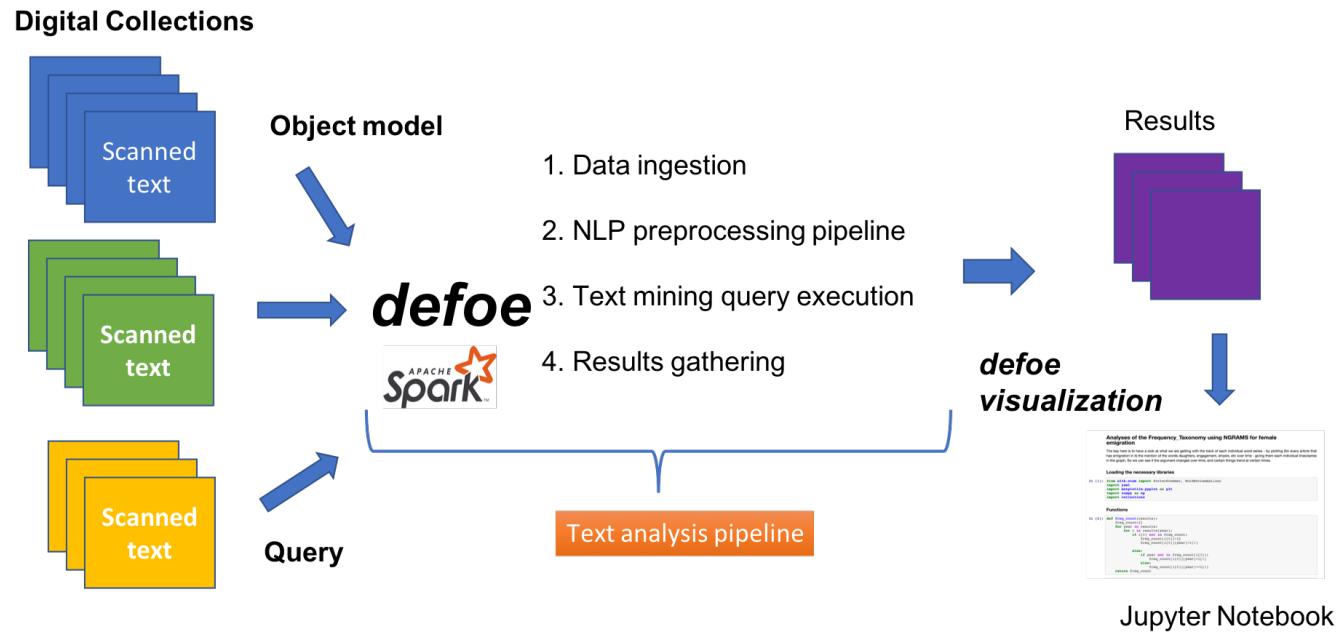
(*) *Frances Wright* (September 6, 1795 – December 13, 1852)

The Journey

- (1) From Semi-Structured EB information (ALTO & METS XML files) →**
Information Extraction → Knowledge Graph → Deep Transfer learning
→ **To Augmented EB-Knowledge Graph with advance AI-method**

- (2) Querying Augment EB Knowledge Graph:**
 - (1) Extracting information already stored in the EB-Knowledge Graph
 - (2) Processing information stored in the EB-Knowledge Graph **in parallel**
→ **Create new data/results**

Phase 1: Text Mining



METS & ALTO-XML

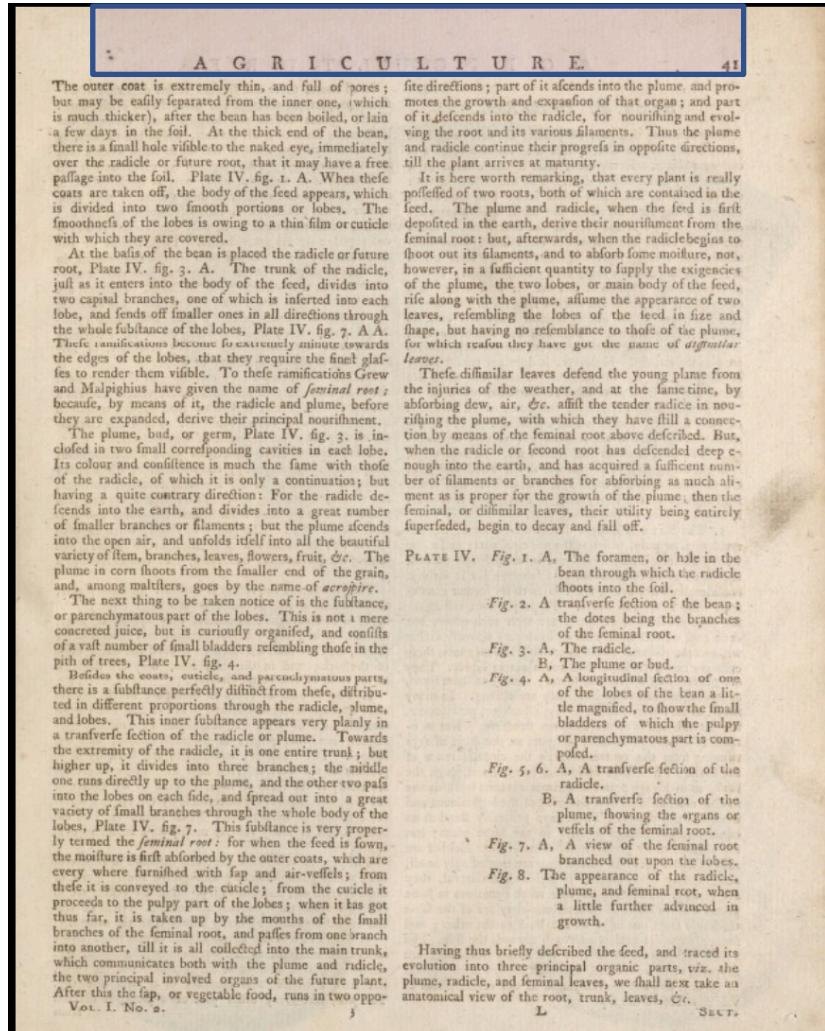
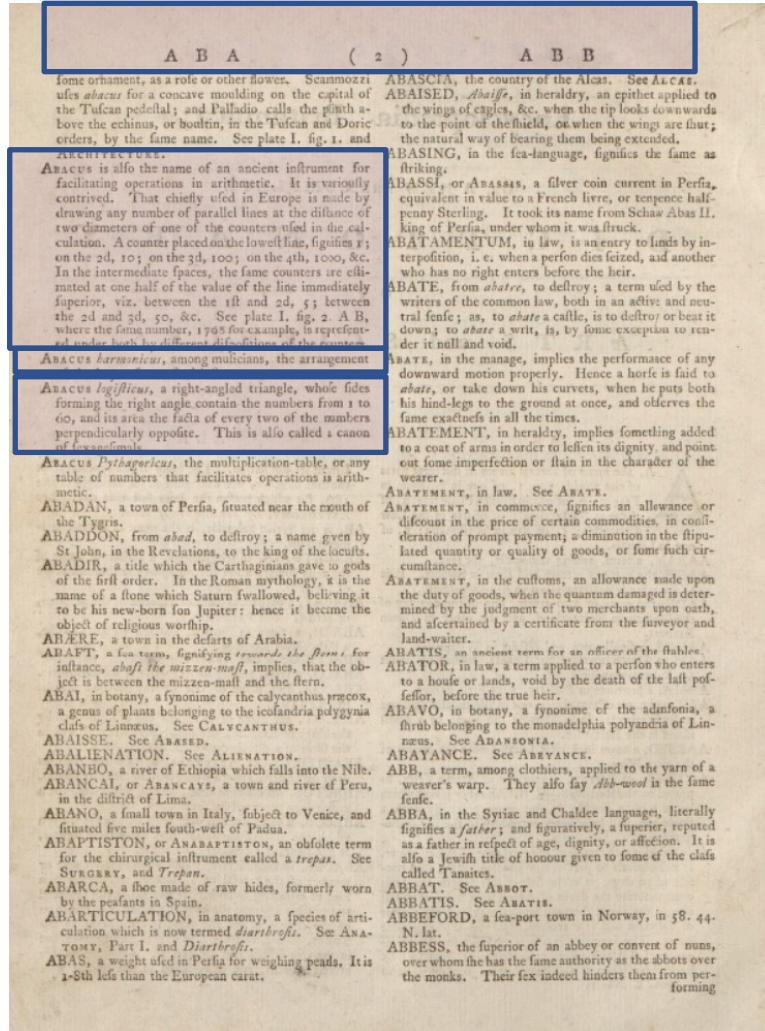
Phase 1: Text Mining

1.1. Improved **defoe Extract Terms query** → It extracts the Eb Terms by page and classify them between **articles** and **topics** (**Terms v.1**)

Extracted Terms Based on Heuristics --`> Pages Layout & Text & Headers → Different heuristics for different EB editions. Using ALTO-XML information.

Phase 1: Text Mining

1. Detecting pages headers from ALTO XML
2. Using headers to classify terms into: Articles & Topics
3. Using ALTO Text for detecting the start of each article:
--> Starting a line with TERM UPPERCASE + “,” .



Articles

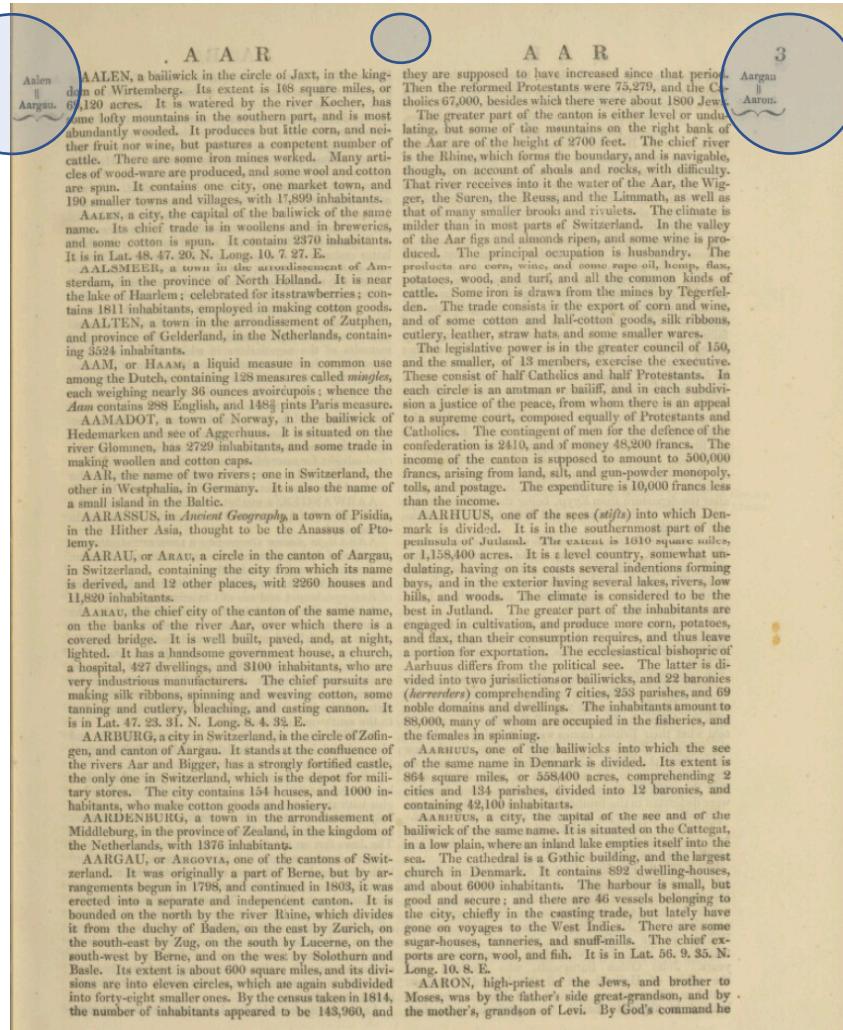
Edition 1 - 1771

Topic

Phase 1: Text Mining

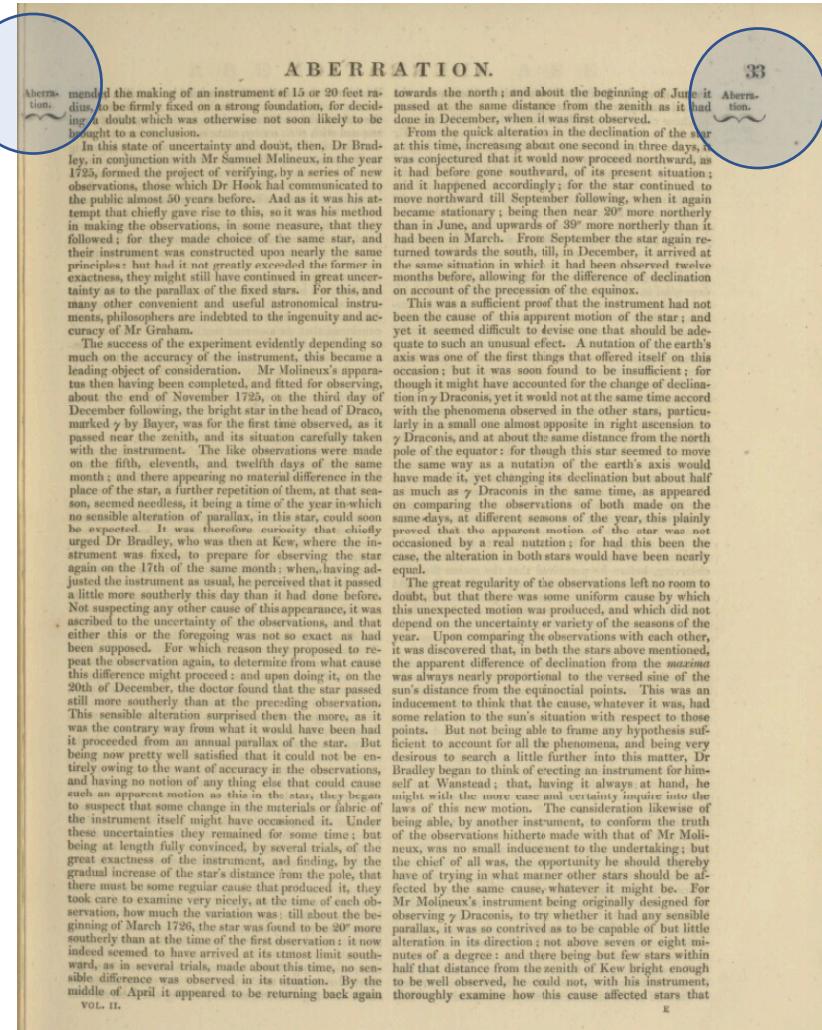
1. Detecting pages headers from ALTO XML
2. Using headers to classify terms into: Articles & Topics
3. Using ALTO Text for detecting the start of each article:

--> Starting a line with TERM UPPERCASE + “,” .



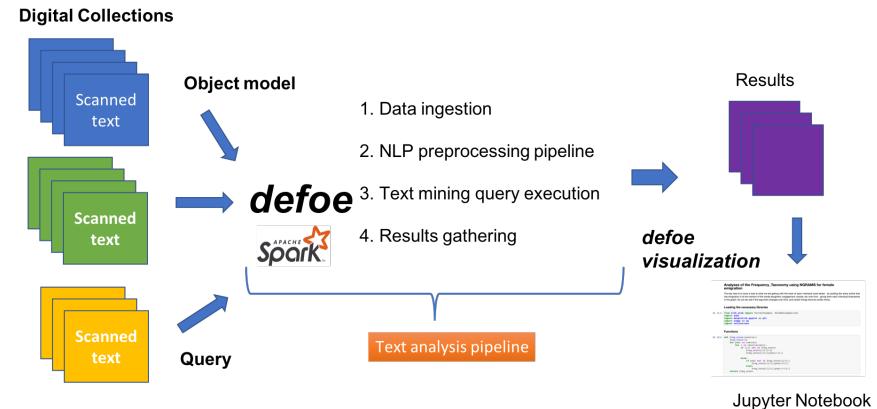
Articles

Edition 7



Topic

Extract EB Terms



term
definition
relatedTerms
header
startsAt
endsAt
numberOfTerms
numberOfWords
numberOfPages
positionPage
typeTerm
editionTitle
editionNum
supplementTitle
supplementsTo
year
place
volumeTitle
volumeNum
letters
part
altoXML
Name: 18, dtype: object

ABACTORES
or ABACTORS, a term for such as carry offer dr...
[]
EBAA
15
15
22
18
832
18
Article
First edition, 1771, Volume 1, A-B
1
[]
1771
Edinburgh
Encyclopaedia Britannica; or, A dictionary of ...
1
A-B
0
144133901/alto/188082904.34.xml

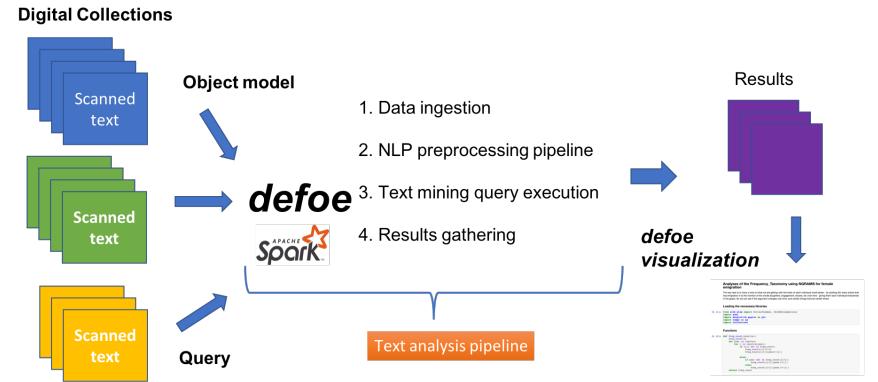
Extracted Term ABACATORES information - Edition 1, 1771, Vol A-B

Phase 1: Text Mining

1.2 Improved defoe **Metadata Extraction query** → It extracts the metadata per Edition and Volume (**Metadata v.1**)

Based on METS information

Extract collection Metadata (METS)



	MMSID	editionTitle	editor	editor_date	genre	language	termsOfAddress	numberOfPages	physicalDescription	place	...	permanentURL	Jupyter Notebook
14	997902543804341	Third edition, Volume 2, ANG-BAR	None	None	encyclopedia	eng	None	922	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149977338	
15	997902543804341	Third edition, Volume 3, BAR-BZO	None	None	encyclopedia	eng	None	856	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149977873	
16	997902543804341	Third edition, Volume 4, CAA-CIC	None	None	encyclopedia	eng	None	842	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149978642	
17	997902543804341	Third edition, Volume 5, CIC-DIA	None	None	encyclopedia	eng	None	858	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149979156	
18	997902543804341	Third edition, Volume 6, DIA-ETH	None	None	encyclopedia	eng	None	850	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149979622	
19	997902543804341	Third edition, Volume 7, ETM-GOA	None	None	encyclopedia	eng	None	882	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149981189	
20	997902543804341	Third edition, Volume 8, GOB-HYD	None	None	encyclopedia	eng	None	832	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149981670	
21	997902543804341	Third edition, Volume 9, Hydrostatics- LES	None	None	encyclopedia	eng	None	872	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149982181	
22	997902543804341	Third edition, Volume 10, LES-MEC	None	None	encyclopedia	eng	None	842	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149982692	
23	997902543804341	Third edition, Volume 11, Medals- Midwifery	None	None	encyclopedia	eng	None	862	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/149983206	
26	997902543804341	Third edition, Volume 1, A-ANG	None	None	encyclopedia	eng	None	894	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/190273291	
27	997902543804341	Third edition, Volume 12, MIE-NEG	None	None	encyclopedia	eng	None	870	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/190273372	
28	997902543804341	Third edition, Volume 13, NEH-PAS	None	None	encyclopedia	eng	None	874	18v.,plates : ill.,maps,music ; 4to	Edinburgh	...	https://digital.nls.uk/191253798	

Metadata of some of the Volumes of Edition 3

Phase 1: Text Mining

1.3. **New Post-processing python scripts (*)** to improve the previous results:

- Re-classification of **Terms v.1** (articles and topics)
- Join terms spitted across pages

We get here: **Terms & Metadata v.2**

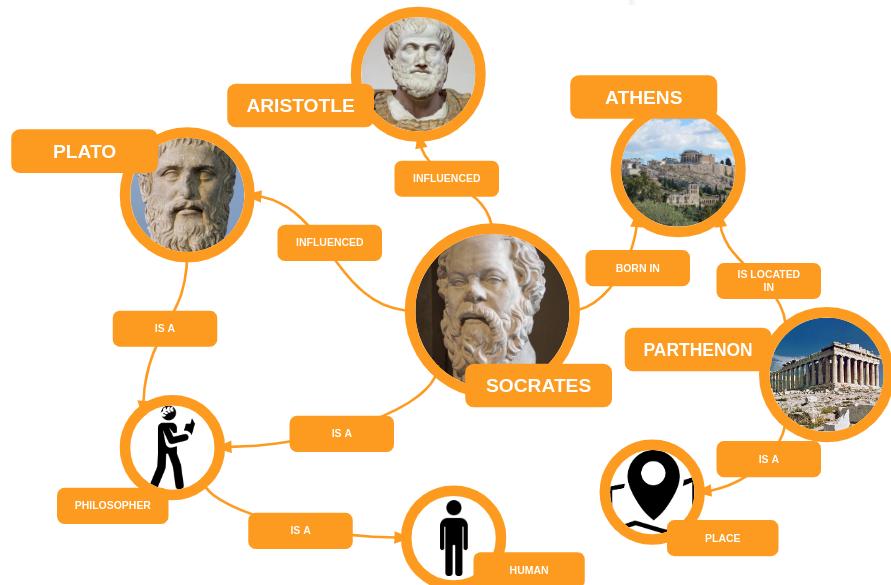
(*) **Also Based on Heuristics --`>** Pages Layout & Text & Headers → different heuristics for different EB editions.

Phase 2: Knowledge Graph

Knowledge Graph: Incorporate human knowledge into intelligent systems, exploiting a semantic graph perspective

- A **knowledge graph** is a specialized graph or network of the things we want to describe and how they are related
- It is a **semantic** model since we want to capture and generate **meaning** with the model

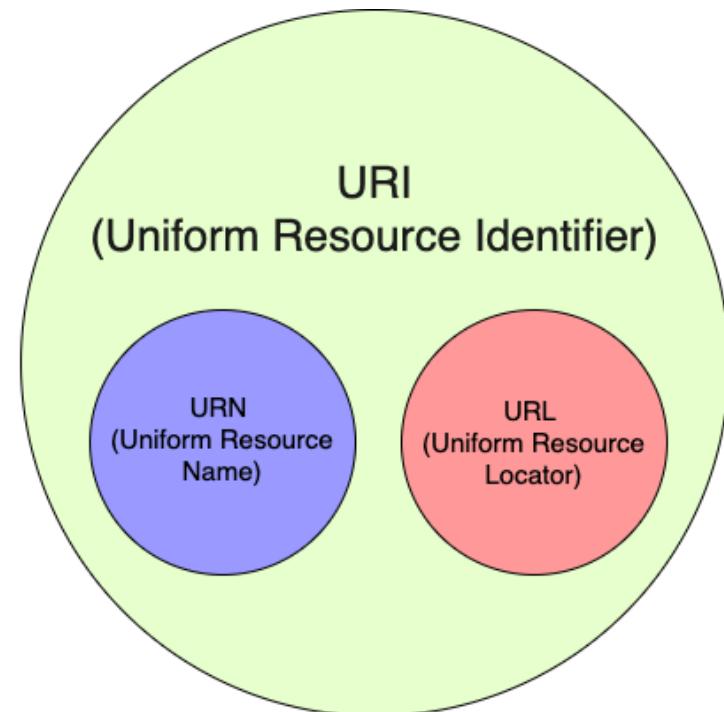
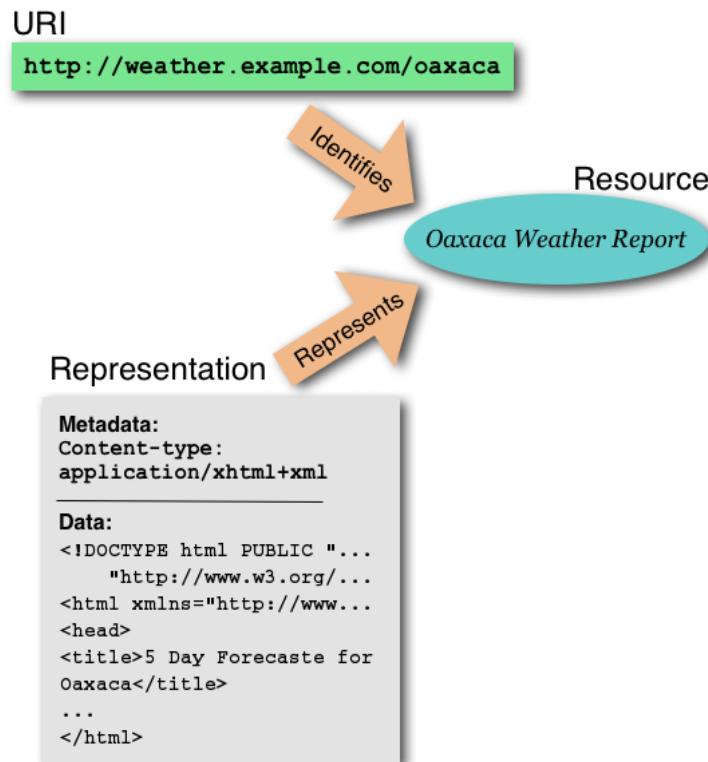
"The application of graph processing and graph DBMSs will grow at 100 percent annually through 2022 to continuously accelerate data preparation and enable more complex and adaptive data science."
– Gartner's *Top 10 Data and Analytics Technology Trends for 2019*



Phase 2: Knowledge Graph

Knowledge Graph: Incorporate human knowledge into intelligent systems, exploiting a semantic graph perspective

URI: A Universal Resource Identifier, is defined to be an ASCII string used to identify “things” on the Knowledge Graph



Phase 2: Knowledge Graph

Knowledge Graph: Incorporate human knowledge into intelligent systems, exploiting a semantic graph perspective

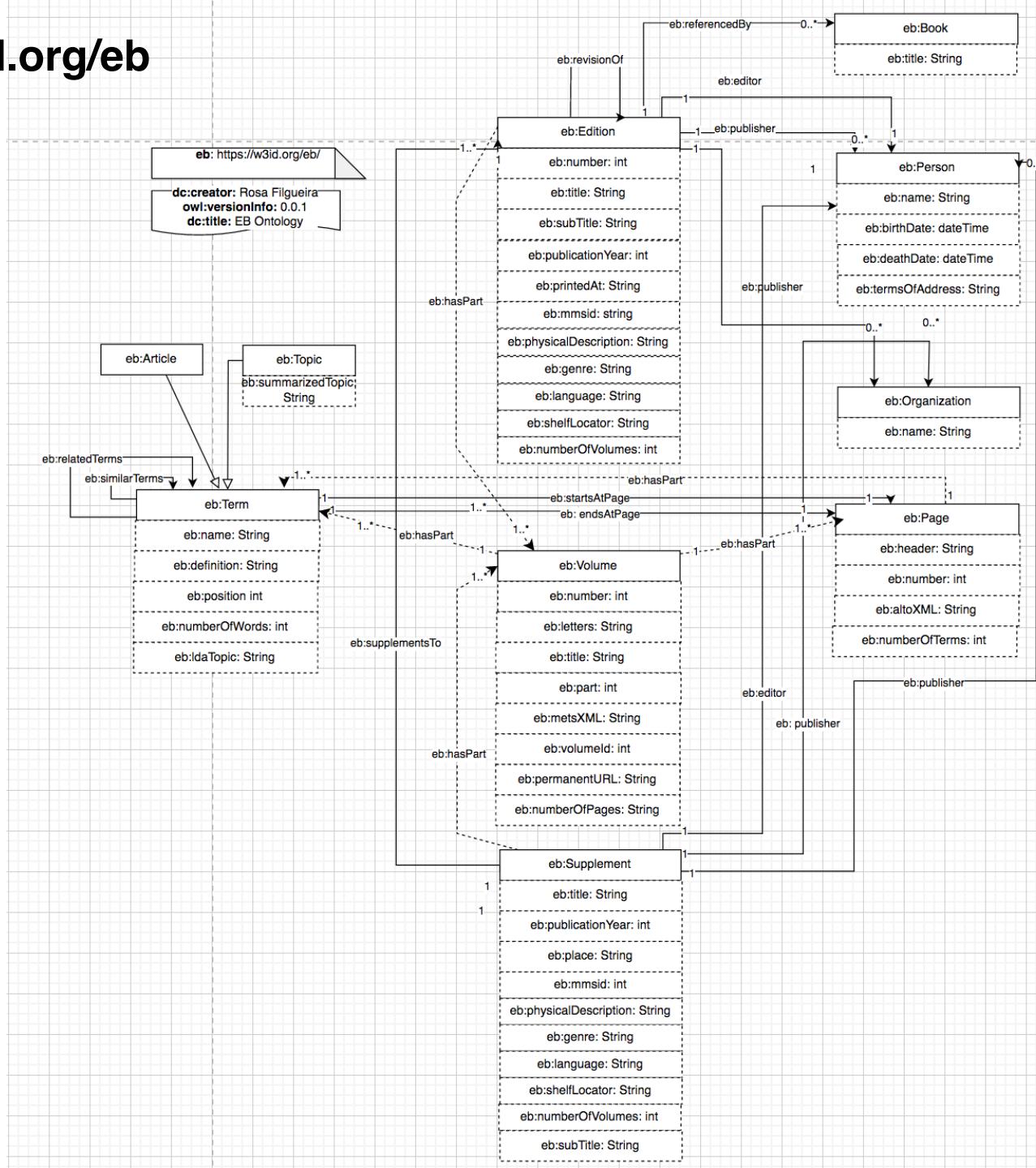
2.1. Create EB Ontology: To explain (give meaning) how our “thing” are related with each other.

In order to create and publish the EB-Ontology I used:

- [diagrams.net](#) : To create an UML with the EB information (classes, properties, relationships, etc.)
- [Chowlk](#) : To convert the UML into an OWL ontology
- [Widoco](#): To publish and create an enriched and customized documentation of the ontology
- [w3id.org](#): To configure my permanent Identifier for EB ontology →
<https://w3id.org/eb/>

EB-Ontology : <https://github.com/francesNLP/EB-ontology>

<https://w3id.org/eb>



Phase 2: Knowledge Graph

Knowledge Graph: Incorporate human knowledge into intelligent systems, exploiting a semantic graph perspective

2.1. Created EB Ontology: The description of this ontology is available online at http: <https://w3id.org/eb/>

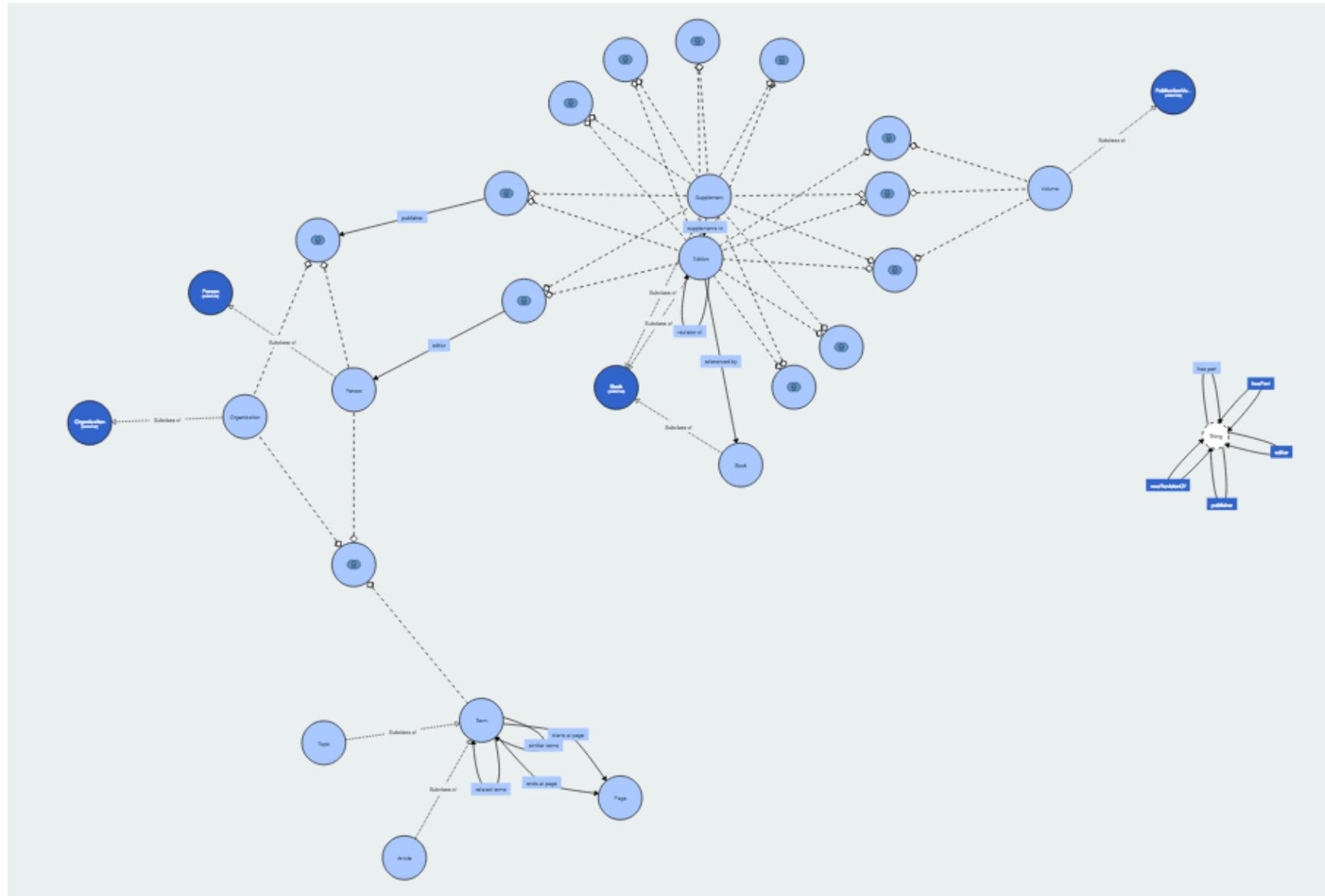
2.2 . Created **EB Knowledge Graph 1.0**: Populated the post-processed information (extracted **Terms & Metadata v.2**) into a RDF triplestore using the EB Ontology

- My RDF EB-data is stored in an Apache Jena FUSEKI SPARQL server – Used this Fuseki-docker image to set up my SPARQL server/end-point

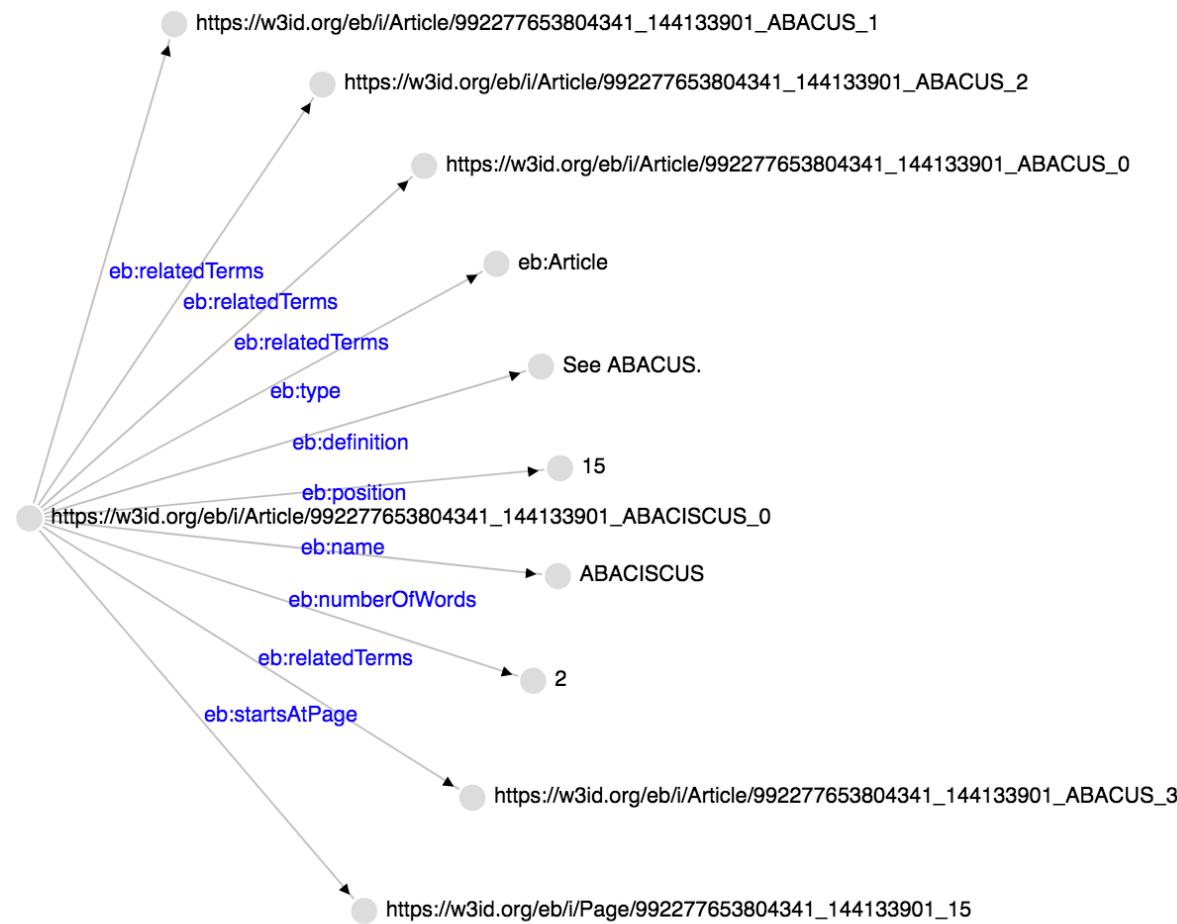
NOTE:

Each Term, Edition, Page, Volume, etc ... is a **Resource in our Knowledge Graph** and has an **URI to identify it**.

EB Knowledge Graph

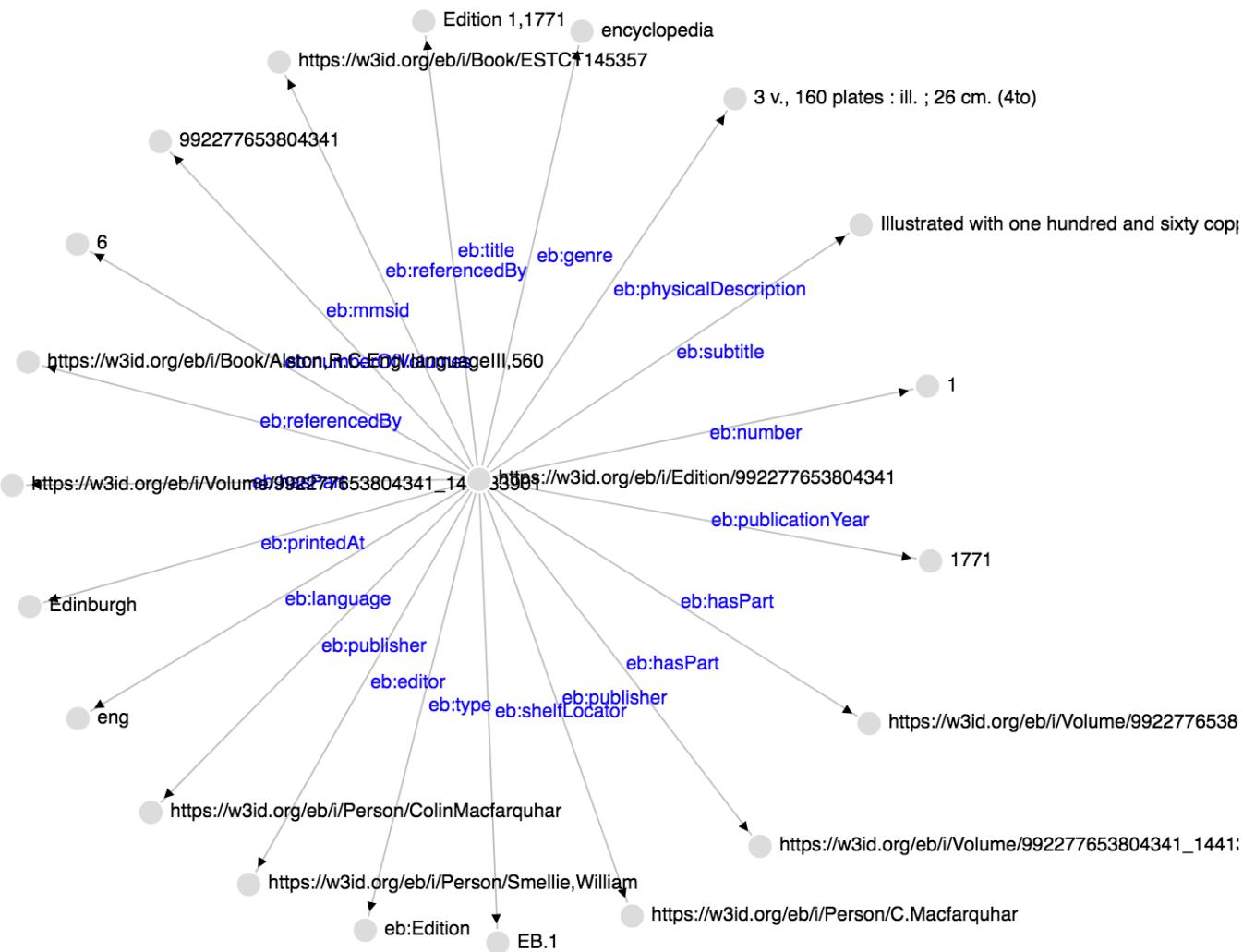


EB Knowledge Graph



Term ABACISCUS, URI: https://w3id.org/eb/i/Article/992277653804341_144133901_ABACUS_0

EB Knowledge Graph



Edition 1, 1771 – URI: <https://w3id.org/eb/i/Edition/992277653804341>

EB Knowledge Graph – Querying our KG

SPARQL is an **RDF query language**—that is, a semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework (RDF) format.

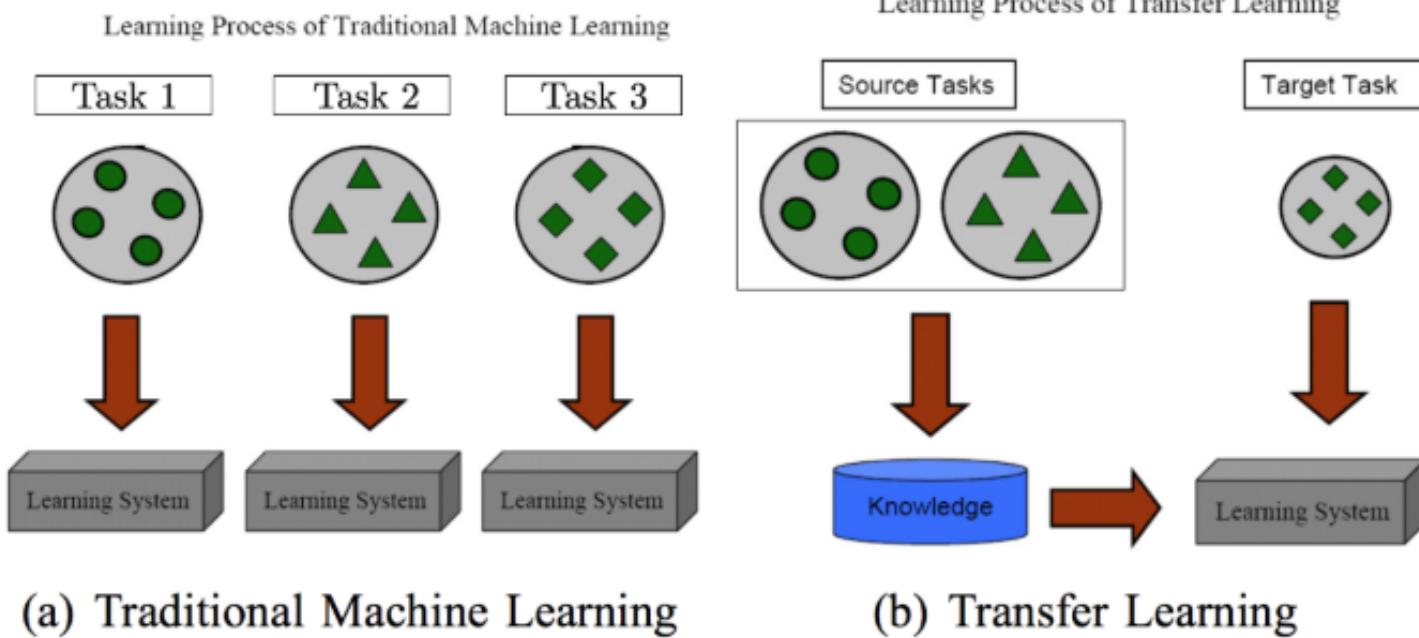
```
2 SELECT ?article ?definition
3 WHERE {
4     ?article a eb:Article .
5     ?article eb:definition ?definition
6     FILTER (CONTAINS(?definition, "Scotland"))
7     FILTER (CONTAINS(?definition, "Glasgow"))
8     OPTIONAL{FILTER CONTAINS(?definition, "Edinburgh") }
9 } LIMIT 10
10
```



Fuseki RDF triplestore

QUERY RESULTS	
Table	Raw Response
Download	
Showing 1 to 8 of 8 entries	
	<input type="text"/> Search: <input type="button"/> Show 50 entries
article	definition
1 https://w3id.org/eb/l/Article/9929192893804340_144850368_PAISLEY_0	"a town of Scotland, in the county of Renfrew, six miles west of Glasgow."
2 https://w3id.org/eb/l/Article/9929192893804340_144850366_ARGYLESHIRE_0	"a county of Scotland, lying westward of Glasgow, and comprehending the countries of Lorn, Cowal, Knapdale, Kintyre, together with the islands Mull, Jura, Iona, & Canna. It gives the title of duke to the noble family of Campbell."
3 https://w3id.org/eb/l/Article/9929192893804340_144850367_INVERARY_0	"a parliament town of Scotland, in the county of Argyle, of which it is the capital, situated in Lochay, forty five miles north-west of Glasgow : W. long. 5° 0', N. lat 36° 28'."
4 https://w3id.org/eb/l/Article/9929192893804340_144850367_HAMILTON_0	"a town of Scotland, in the county of Clydesdale, situated on the river Clyde, eleven miles south-east of Glasgow : W. long. 3° 0', N. lat. 55° 0"

Phase 3: Augmented Knowledge Graph with Deep Transfer Learning



Phase 3: Augmented Knowledge Graph with Deep Transfer Learning



Transformers

build passing license [Apache-2.0](#) website [online](#) release [v2.0.0](#)

State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0

😊 Transformers provides thousands of pretrained models to perform tasks on texts such as classification, information extraction, question answering, summarization, translation, text generation, etc in 100+ languages. Its aim is to make cutting-edge NLP easier to use for everyone.

😊 Transformers provides APIs to quickly download and use those pretrained models on a given text, fine-tune them on your own datasets then share them with the community on our [model hub](#). At the same time, each python module defining an architecture can be used as a standalone and modified to enable quick research experiments.

😊 Transformers is backed by the two most popular deep learning libraries, [PyTorch](#) and [TensorFlow](#), with a seamless integration between them, allowing you to train your models with one then load it for inference with the other.

Phase 3: Augmented Knowledge Graph with Deep Transfer Learning



build passing license Apache-2.0 website online release v2.0.0

A High-Level Look

Let's begin by looking at the model as a single black box. In a machine translation application, it would take a sentence in one language, and output its translation in another.



Phase 3: Augmented Knowledge Graph with Deep Transferring Learning

EB Knowledge Graph 2.0: previous info + storing the result of applying different deep learning transformers analyses:

- **sentiment analyses:** Classifying text between positive and negative
 - transformer: [*siebert/sentiment-roberta-large-english*](#)
- **topic modelling:** Clustering terms into topics
 - transformer: [*all-mpnet-base-v2*](#)
- **term similarity:** Comparing text & semantic similarity
 - transformer: [*all-mpnet-base-v2*](#)
- **spelling checking:** Finding misspelling/ocr errors and fixing them
 - transformer: [*neuspell*](#) + *ElmoslstmChecker*
- **term evolution:** Checking how a term has changed over the years
 - transformer: [*all-mpnet-base-v2*](#)
- **summarization:** Summarizing the text of a topic term (XLNET)
 - transformer: [*XLNet*](#)

Phase 3: Augmented Knowledge Graph with Deep Transferring Learning

EB Knowledge Graph 2.0: previous info + storing the result of applying different deep learning transformers analyses:

- **sentiment analyses:** Classifying text between positive and negative
- **topic modelling:** Clustering terms into topics
- **term similarity:** Comparing text & semantic similarity
- **spelling checking:** Finding misspelling/ocr errors and fixing them
- **term evolution:** Checking how a term has changed over the years
- **summarization:** Summarizing the text of a topic term

Example: Spelling Checking → *Lewis* Term

Original Definition

the mort northerly of any of the w eftern islands of Scotland, lying in 8\u00b0 odd minutes W. long, and between 58\u00b0 and 59 0 odd minutes N. lat.



Cleaned Definition

the most northerly of any of the w eastern islands of Scotland , lying in 8 and odd minutes W. long , and between 58 and and 59 0 odd minutes N. land .



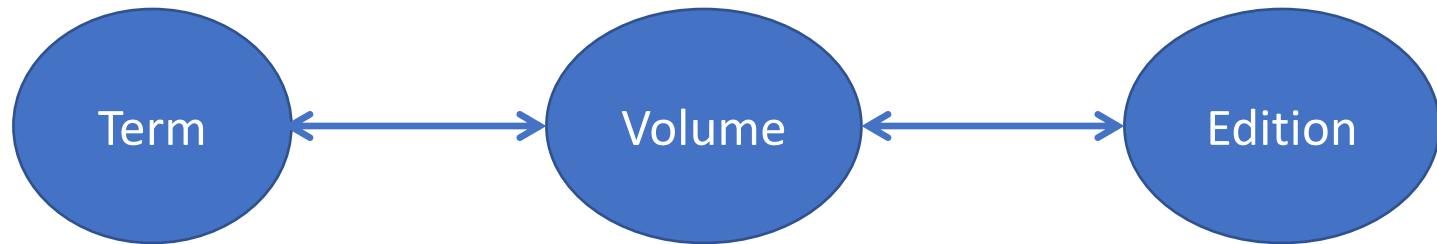
Compute Difference

the morst northerly of any of the w eftern islands of Scotland, lying in 8\u00b0 odd minutes W. long, and between 58\u00b0 and 59 0 odd minutes N. latnd .

Phase 3: Augmented Knowledge Graph with Deep Transferring Learning

Two types of “queries” against the EB Knowledge Graph 2.0 :

- **Type 1: Extracting information from the EB Knowledge → SPARQL → We “just” navigate through the KG to get the desired information.**

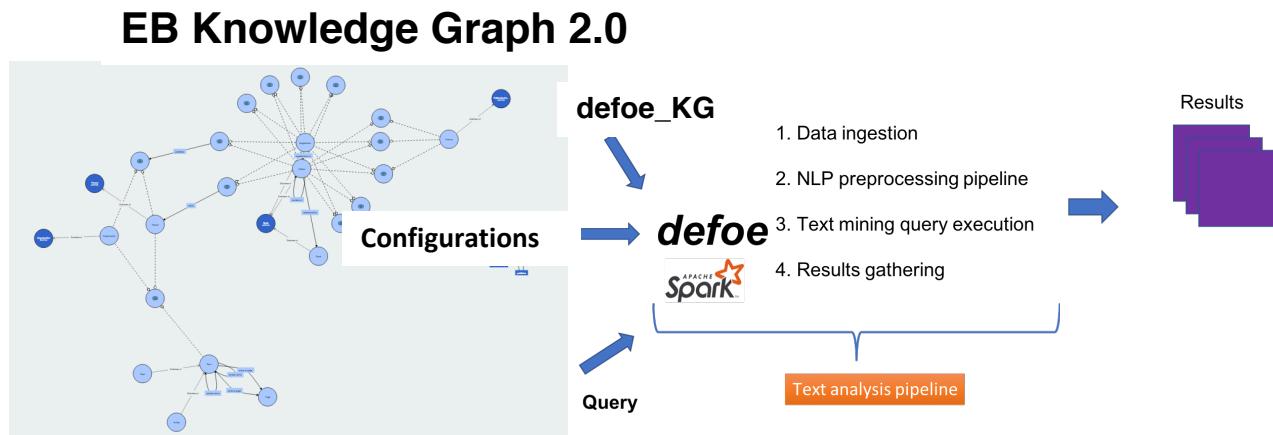


Example: Given a **Term** (e.g., *Edinburgh*), I can get the **Edition** Information (e.g., *Edition Title*)

- **Type 2: Processing information from the EB Knowledge → defoe → We are going to process further the definitions from the selected terms.**
 - But for doing this we needed to do some work on defoe first → Phase 4

Phase 4: Defoe and Knowledge Graph

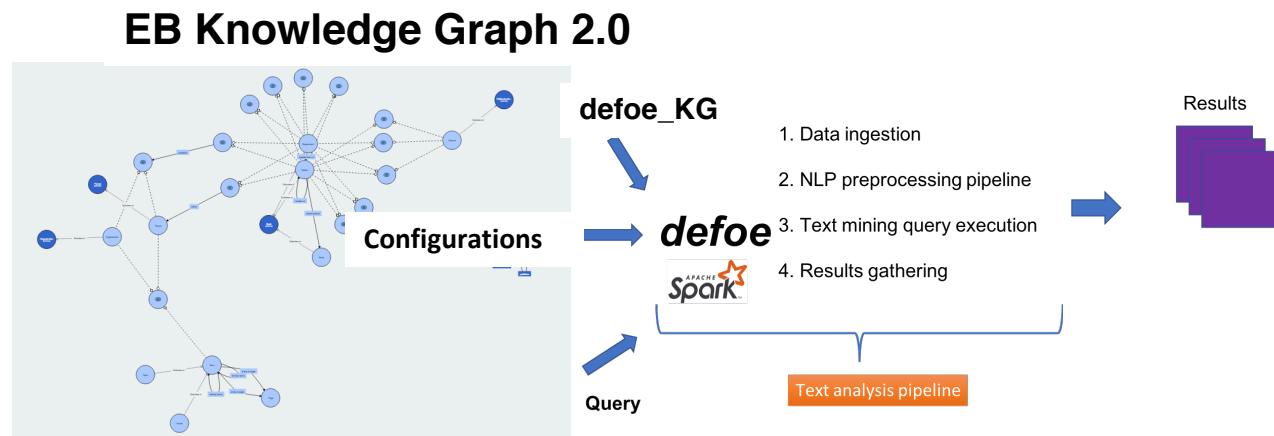
- 4.1. Created a new **KG defoe connector** (based in SPARQL) to run defoe queries using the EB Knowledge Graph as a source of data → **defoe_KG**
- 4.2. Improved **defoe queries to be fully configurable**: different filtering options, target, lexicon, etc.



Phase 4: Defoe and Knowledge Graph

4.3 defoe text mining queries:

- frequency keysearch: **Count number of terms** or times in which appear keywords or keysentences and group by years. Several filtering options.
- term fulltext keysearch: **Extract terms definitions** in which appear keywords or keysentences and group by years. Several filtering options.
- term snippet keysearch: **Extract snippet of definition** in which appear keywords or keysentences and group by years. Several filtering options, including the snippet size.
- publication normalization: **Extract number of documents, pages, words** per year.
- uris keysearch: **Extract uris of terms** in which appear keywords or keysentences and group by years. Several filtering options.
- geoparser terms: **Geoparsing the term definition** in which appear keywords or keysentences and group by years. Several filtering options.



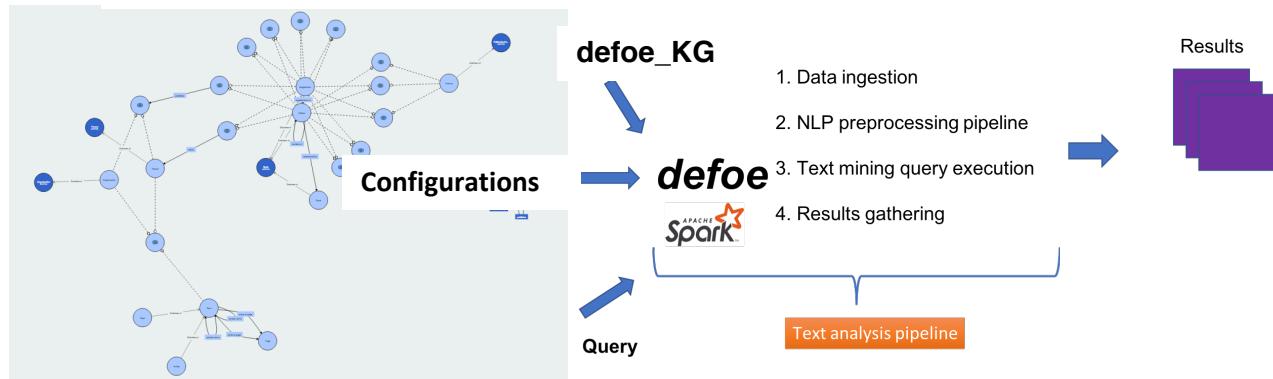
Phase 4: Defoe and Knowledge Graph

4.4 Created **new defoe ML/NLP queries** (using also deep learning transformers)

- sentiment analysis: **calculate the sentiment** analyses of selected terms.
- topic modelling: **calculate the topic** of selected terms. Filtering options.
- spelling checker: **check the spelling** of selected terms. Filtering options.

(TO-DO)

EB Knowledge Graph 2.0



Phase 5: Flask Web-Application

5.1 Web-User Interface:

- Term Search
- Term Similarity
- Topic Modelling
- Spelling Checker
- Defoe queries
- EB Details
- Knowledge Graph Resources Visualizations

5.2 Flask + JQuery + JavaScript + HTML + CSS + Web-services + Plotly

Summary

1. EB ALTO & METS + **defoe** → Extracted **Terms & Metadata (v.1)**
2. Extracted terms and metadata (v.1) + **postprocessing scripts**:
 1. Final set of **Terms & Metadata (v.2)**
 2. **EB ontology**
3. Terms & Metadata (v.2) + EB ontology → **EB Knowledge Graph 1.0**
4. EB Knowledge Graph 1.0 + Transformers → **EB Knowledge Graph 2.0**
5. EB Knowledge Graph 2.0 + defoe → **defoe_KG**
6. defoe_KG + **configurable defoe queries**
7. EB Knowledge Graph 2.0 + defoe_KG + defoe queries (text mining/ NLP)
8. EB Knowledge Graph Queries:
 1. **Type 1: SPARQL queries:** Extracting the EB KG data
 2. **Type 2: defoe queries:** Processing the EB KG data
9. Flask-web app

Interacting with the EB Knowledge Graph

Term Search Term Similarity Topic Modelling Spell Checker Term Evolution Defoe Queries EB Details Visualization of Resources

Exploring the Encyclopaedia Britannica (1768-1860)

Term Search

Enter the **term** that you would like to search for. In case that the **Term Type is a Topic**, only the **summary** of the definition is displayed. If no term is introduced, it will search for the first term in the Encyclopaedia.

Results for **EDINBURGH**.

Note that if you click over an **URI** in this table, it will take you to the **Visualization of Resources page**. However, if you instead click over a **related term**, it will conduct a **term search**, showing all the searching results for that term. And if you click over a **topic model** if will take you to the **Topic Modelling page**, listing all the terms belonging to that particular topic model.

displaying 1 - 2 records in total 2

URI	Year	Edition	Volume	Start Page	End Page	Term Type	Definition/Summary	Related Terms	Topic Modelling	Sentiment_Score	Advanced Options
https://w3id.org/eb/i/Article/992277653804341_144133902_EDINBURGH_0	1771	1	2	408	409	Article	the capital city of the kingdom of Scotland, situated W. long. 3° 0', and N. lat. 56° 0' 0". We ffiall... More		15_scotland_county_edinburgh_firth	LABEL_0_0.87	Spell Checker Term Similarity Term Evolution
https://w3id.org/eb/i/Article/9929192893804340_144850367_EDINBURGH_0	1773	1	2	414	415	Article	the capital city of the kingdom of Scotland, situated W. long. 3° 0', and N. lat. 56° 0' 0". We (hall ... More		15_scotland_county_edinburgh_firth	LABEL_0_0.89	Spell Checker Term Similarity Term Evolution

FLASK Web-Application

Interacting with the EB Knowledge Graph

Type 1

Term Search

Term Similarity

Topic Modelling

Spell Checker

Term Evolution

Type 2

Defoe Queries

Type 1

EB Details

Visualization of Resources

Exploring the Encyclopaedia Britannica (1768-1860)

Term Search

Enter the **term** that you would like to search for. In case that the **Term Type is a Topic**, only the **summary** of the definition is displayed. If no term is introduced, it will search for the first term in the Encyclopaedia.

Results for **EDINBURGH**.

Note that if you click over an **URI** in this table, it will take you to the **Visualization of Resources page**. However, if you instead click over a **related term**, it will conduct a **term search**, showing all the searching results for that term. And if you click over a **topic model** it will take you to the **Topic Modelling page**, listing all the terms belonging to that particular topic model.

displaying 1 - 2 records in total 2

URI	Year	Edition	Volume	Start Page	End Page	Term Type	Definition/Summary	Related Terms	Topic Modelling	Sentiment_Score	Advanced Options
https://w3id.org/eb/i/Article/992277653804341_144133902_EDINBURGH_0	1771	1	2	408	409	Article	the capital city of the kingdom of Scotland, situated W. long. 3° 0', and N. lat. 56° 0' 0". We ffiall... More		15_scotland_county_edinburgh_firth	LABEL_0_0.87	Spell Checker Term Similarity Term Evolution
https://w3id.org/eb/i/Article/9929192893804340_144850367_EDINBURGH_0	1773	1	2	414	415	Article	the capital city of the kingdom of Scotland, situated W. long. 3° 0', and N. lat. 56° 0' 0". We (hall ... More		15_scotland_county_edinburgh_firth	LABEL_0_0.89	Spell Checker Term Similarity Term Evolution

Note: Two types of EB-KG queries

- **Type 1:** Extracting information from EB-KG – Navigating across the KG
- **Type 2:** Processing information from EB-KG -- Processing in parallel further the Term's definitions (stored in KG).

frances: Architecture

NSL Data Foundry Collections

