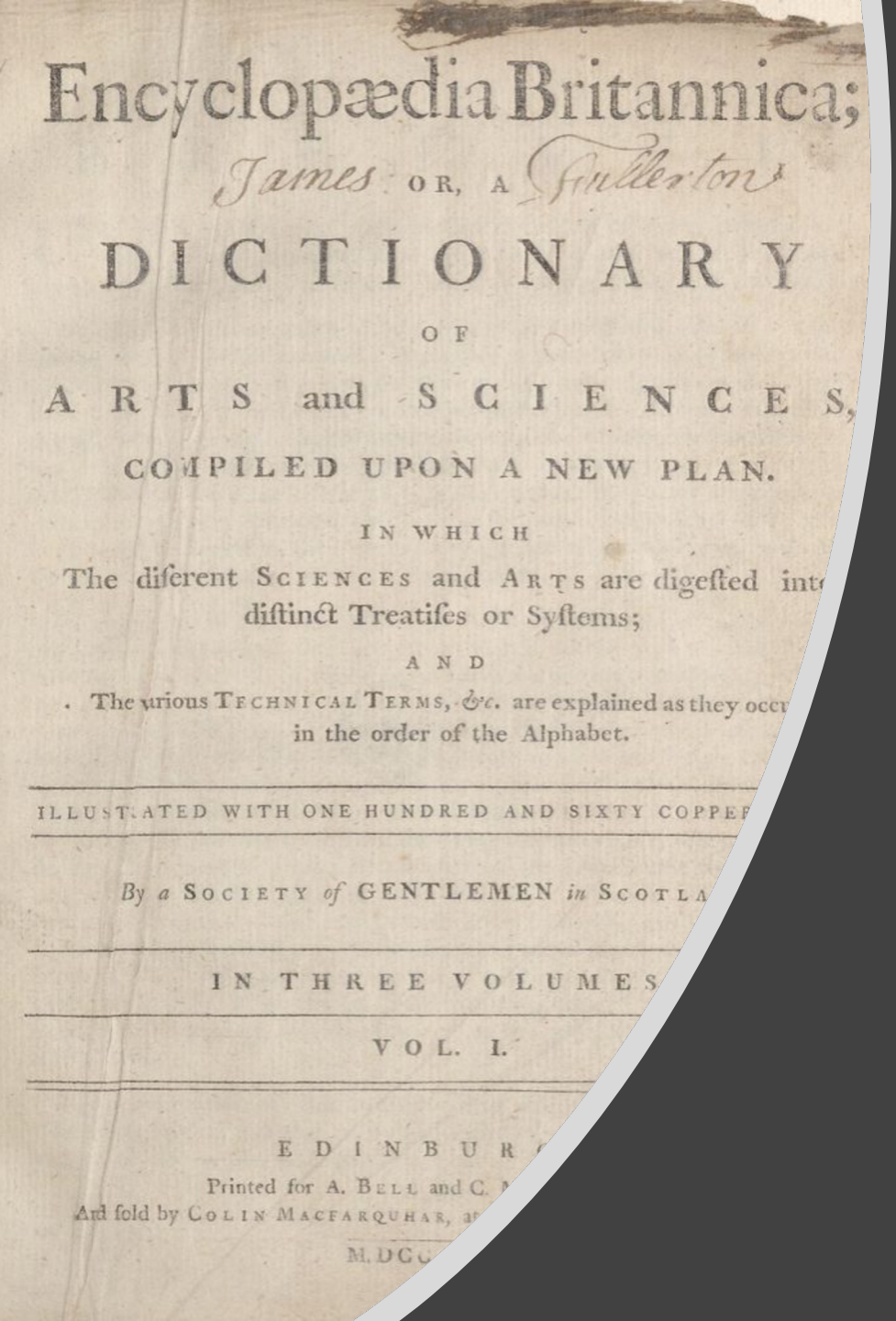


Frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

Dr. Rosa Filgueira,
Lecturer at the School of Computer Science,
University of St Andrews,
Email: rf208@st-Andrews.ac.uk

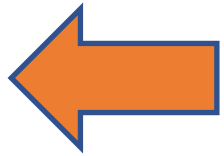
Background – Research Interest

St Andrews / CS

Scalability, reliability, and performance of data science applications

- Data/task workflows & programming abstractions
- Automatic and portable parallelisation techniques
- Workload strategies that automatically adapt to applications at runtime
- Dynamic optimization techniques for stream-based workflows/frameworks
- Techniques for dealing with failures at the application and the system levels
- Exploiting heterogenous platforms for data-intensive applications

Unlocking Historical Digital Text Collections through Advanced AI methods and Parallel Techniques



- Large scale text mining facilities in digital humanities
- Hiding system complexity
- Ease of use and consistency for complex text-mining analysis
- Techniques for making digital text searchable, and analysable
- Complex text analysis queries automatically and at scale, and for visualizing results
- Advanced AI-techniques to extract knowledge from digital text collections.

Intelligent advisory systems for scientific software

- Software feature extraction
- Static code analysis techniques
- Semantic code search algorithms; automatic code classification/tagging methods;
- Automatic pipeline/workflow composition and orchestration;
- -Automated serverless computing architectures

frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

St Andrews / CS

Motivation

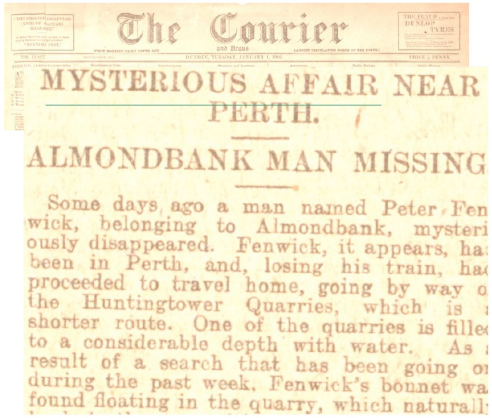
- Large digital collections been available for research
- Hunger for large scale text mining facilities and for:
  - HPC/Cloud environments
  - Analytic frameworks to create applications

Collaborations ~ 6 years:

- *Melisa Terras, Lisa Otty* – CDCS UoE
- *Bea Alex, Sarah van Eyndhoven, Lisa Gotthard* UoE
- *Sarah Ames*, NLS
- *LwM project*- Alan Turing

Challenges

- Several large digital collections (semi-structured data)
- Different levels of quality of data – OCR
- Data with different physical representations and XML schemas
- Difficult to mine, search, analyse, compare



The Courier

MYSTERIOUS AFFAIR NEAR PERTH.

ALMONDBANK MAN MISSING

Some days ago a man named Peter Fen wick, belonging to Almondbank, mysteri ously disappeared. Fenwick, it appears, ha been in Perth, and, losing his train, ha proceeded to travel home, going by way o the Huntingtower Quarries, which is shorter route. One of the quarries is fille to a considerable depth with water. As result of a search that has been going o during the past week, Fenwick's bonnet wa found floating in the quarry, which natural

```
...
<text.title>
  <pg pgref="5" clipref="1"
      pos="4069,3036,4949,3154"/>
  <p>
    <wd pos="4069,3036,4949,3154">MYSTERIOUS AFFAIR
NEAR PERTH.</wd>
  </p>
</text.title>
<text.cr>
  <pg pgref="5" clipref="1"
      pos="4039,3191,4987,4235"/>
  <p>
    <wd pos="4041,3192,4496,3241">ALMONDBANK</wd>
    <wd pos="4523,3200,4663,3246">MAN</wd>
    <wd pos="4696,3198,4976,3250">MISSING.</wd>
    <wd pos="4085,3290,4189,3323">Some</wd>
    <wd pos="4214,3290,4312,3329">days,</wd>
    ...
```

frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

St Andrews / CS

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Using the Library | Digital resources | Catalogues | Shop | News & events | Contact

Join / Log in

Home › Using the Library › Academic research › Fellowships › Digital scholarship

The National Librarian's Research Fellowship in Digital Scholarship

Fellowships

Graham Brown

Digital scholarship

2021-2022 National Library of Scotland Digital Scholarship Fellow

Dr Rosa Filgueira is an Assistant Professor with the School of Mathematical and Computer Sciences at Heriot-Watt University.

Her research background includes high-performance computing, data streaming, data-intensive computing, and large-scale distributed systems.

Rosa has experience handling and mining large digital textual collections. She has worked on several digital humanities and data science projects, co-developing and co-designing text-mining applications at scale.

Dr Rose Filgueira

AI toolbox project

This project will explore new ways to unlock the full value of the National Library of Scotland's Data Foundry collections by building a new AI toolbox called 'frances' and a web user interface that allows researchers to interact with it.

It would help historians (and other users) to extract complex knowledge from digital collections in a fast and transparent manner without having to be an expert data scientist. The toolbox is named after Frances Wright (September 6, 1795 – December 13, 1852), a Scottish-born lecturer, writer,

*Frances Wright (September 6, 1795 – December 13, 1852), was a Scottish-born lecturer, writer, freethinker, feminist, utopian socialist, abolitionist, social reformer, and Epicurean philosopher

frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

St Andrews / CS

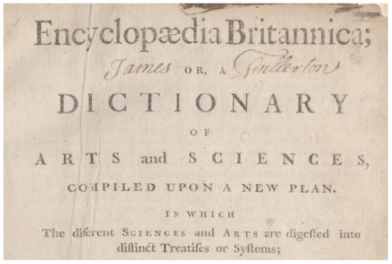
National Library of Scotland
Leabharlann Nàiseanta na h-Alba

HOME ABOUT DATA TOOLS PROJECTS CONTACT

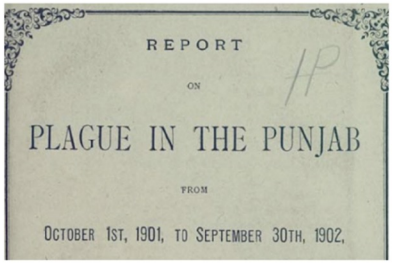
Digitised collections

Download the ALTO, METS, image and plain text files for our digitised collections.

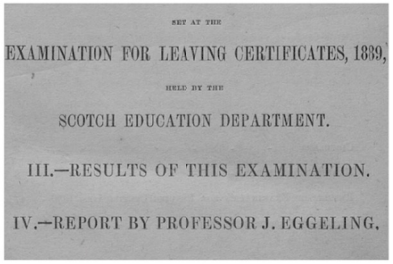
8 Editions:
1768 -1860



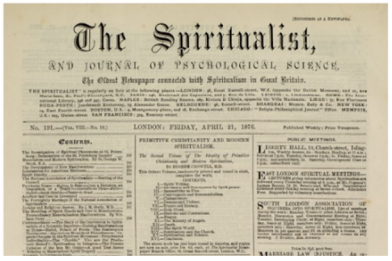
Encyclopaedia Britannica, 1768-1860

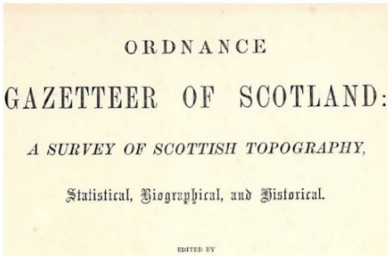


A Medical History of British India



Scottish School Exam Papers, 1888-1963



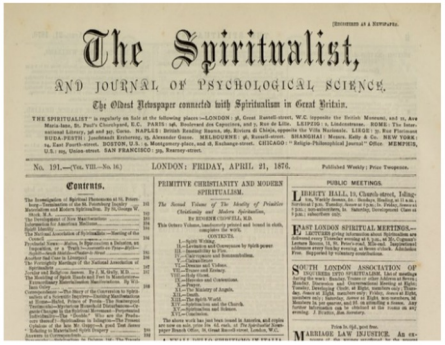


frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

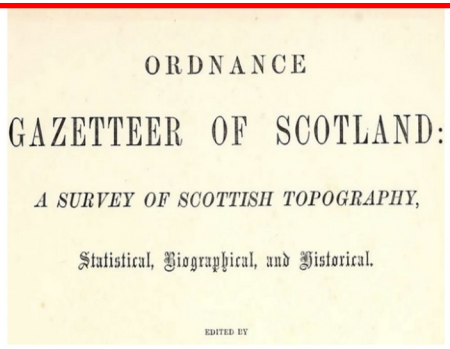
St Andrews / CS

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

HOME ABOUT DATA TOOLS PROJECTS CONTACT



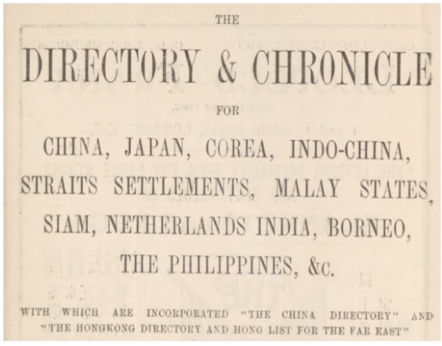
The Spiritualist



Gazetteers of Scotland

Ladies' Edinburgh Debating Society

Chapbooks printed in Scotland



Asian Directories and Chronicles

Britain and UK Handbooks

frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

St Andrews / CS

Goal: Provide **abstractions** to a variety of ML/NLP techniques

→ *Extract complex knowledge without being an expert data scientist*

- Train and use text embedding models
- Employ topic mining, sentiment analysis, text summarization
- Build knowledge graph(s) visualizing the results
- Text Mining Parallel Platform (based in Apache Spark)
- Automatic Visualization of Results

New suite of ML functionalities that can be used to analyse any other Data Foundry collection

Applying advanced IE/AI/NLP to historical digital collections

St Andrews / CS

The Journey

Phase 1: Information Extraction
Phase 2: Ontologies and Knowledge Graphs
Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning
Phase 4: Defoe and Knowledge Graphs
Phase 5: React-Flask Web Platform
Phase 6: Case Studies

Applying advanced IE/AI/NLP to historical digital collections

St Andrews / CS

The Journey

- **Phase 1: Information Extraction**
  - **1.1 Encyclopedia Britannica**
  - **1.2 Others**
- Phase 2: Ontology and Knowledge Graphs
- Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning
- Phase 4: Defoe and Knowledge Graphs
- Phase 5: React-Flask Web Platform
- Phase 6: Case Studies

Phase 1.1: Information Extraction - EB

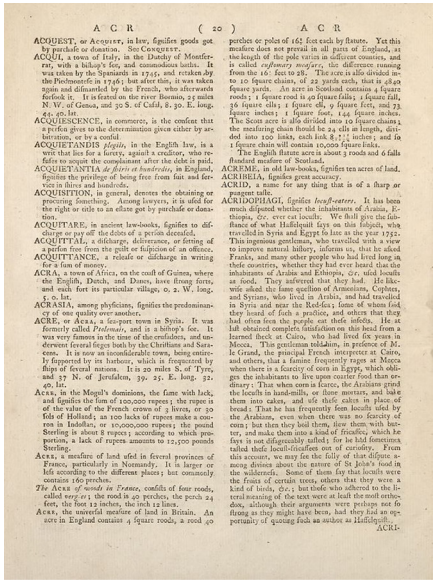
Detect, Classify and Extract all EB terms across editions

St Andrews / CS

- Two types of Terms:
  - **Articles:** 1 or 2 paragraphs describing a term ~ entry in a dictionary
  - **Topics**: Several pages describing a term

- Extract Metadata from all EB editions and volumes.

How have we done it ? By using *defoe* to analyze METS and ALTO XML files



A C R (20) A C R

ACQUEST, or ACQUISE, in law, signifies goods got by purchase or donation. See CONQUEST.

ACQUI, a town of Italy, in the Dutchy of Montferrat, with a bishop's see, and commodious baths. It was taken by the Spaniards in 1745, and retaken by the Piedmontese in 1746: but after this, it was taken again and dismantled by the French, who afterwards forlook it. It is seated on the river Bormio, 25 miles N. W. of Genoa, and 30 S. of Casal, E. 30. E. long. 44. 40. lat.

ACQUIESCENCE, in commerce, is the consent that a perfon gives to the determination given either by arbitration, or by a confol.

ACQUIETANDIS *plegiis*, in the Engliſh law, is a writ that lies for a ſurety, againſt a creditor, who refuſes to acquit the complainant after the debt is paid.

ACQUIETANTIA *de fideiis et hundredis*, in England, ſignifies the privilege of being free from ſuit and ſervice in ſhires and hundreds.

ACQUISITION, in general, denotes the obtaining or procuring ſomething. Among lawyers, it is uſed for the right or title to an eſtate got by purchaſe or donation.

ACQUITTARE, in ancient law-books, ſignifies to diſcharge or pay off the debts of a perſon deceaſed.

ACQUITTAL, a diſcharge, deliverance, or ſetting of a perſon free from the guilt or ſuſpicion of an offence.

ACQUITTANCE, a releaſe or diſcharge in writing for a ſum of money.

ACRA, a town of Africa, on the coaſt of Guinea, where the Engliſh, Dutch, and Danes, have ſtrong forts, and each fort its particular village. 0. 2. W. long. 5. 0. lat.

ACRASIA, among phyſicians, ſignifies the predominancy of one quality over another.

ACRE, or ACRA, a ſea-port town in Syria. It was formerly called *Ptolemais*, and is a biſhop's ſee. It was very famous in the time of the croſaders, and underwent ſeveral ſieges both by the Chriſtians and Saracens. It is now an inconſiderable town, being entirely ſupported by its harbour, which is frequented by ſhips of ſeveral nations. It is 20 miles S. of Tyre, and 37 N. of Jeruſalem, 39. 25. E. long. 32. 40. lat.

ACRE, in the Mogul's dominions, the ſame with lack, and ſignifies the ſum of 100,000 rupees; the rupee is of the value of the French crown of 3 livres, or 30 ſols of Holland; an 100 lacks of rupees make a crore in Indoſtan, or 10,000,000 rupees; the pound Sterling is about 8 rupees; according to which proportion, a lack of rupees amounts to 12,500 pounds Sterling.

ACRE, a meaſure of land uſed in ſeveral provinces of France, particularly in Normandy. It is larger or leſs according to the different places; but commonly contains 160 perches.

The ACRE *of woods in France*, conſiſts of four roods, called *perches*; the rood is 40 perches, the perch 24 feet, the foot 12 inches, the inch 12 lines.

ACRE, the univerſal meaſure of land in Britain. An acre in England contains 4 ſquare roods, a rood 40 perches or poles of 16½ feet each by ſtatute. Yet this meaſure does not prevail in all parts of England, as the length of the pole varies in different counties, and is called *cuſtomary meaſure*, the difference running from the 16½ feet to 28. The acre is alſo divided into 10 ſquare chains, of 22 yards each, that is 484 ſquare yards. An acre in Scotland contains 4 ſquare roods; 1 ſquare rood is 40 ſquare falls; 1 square fall, 36 ſquare ells; 1 ſquare ell, 9 ſquare feet, and 72 ſquare inches; 1 ſquare foot, 144 ſquare inches. The Scots acre is alſo divided into 10 ſquare chains; the meaſuring chain ſhould be 24 ells in length, divided into 100 links, each link 8 4/5 inches; and ſo 1 ſquare chain will contain 10,000 ſquare links.

The Engliſh ſtatute acre is about 3 roods and 6 falls ſtandard meaſure of Scotland.

ACREME, in old law-books, ſignifies ten acres of land.

ACRIBELA, ſignifies great accuracy.

ACRID, a name for any thing that is of a ſharp or pungent taſte.

ACRIDOPHAGI, *ſignifies locuſt-eaters*. It has been much diſputed whether the inhabitants of Arabia, Ethiopia, &c. ever eat locuſts. We ſhall give the ſubſtance of what Haſſelquiſt ſays on this ſubject, who travelled in Syria and Egypt to lay as the year 1752.

This ingenious gentleman, who travelled with a view to improve natural hiſtory, informs us, that he aſked Franks, and many other people who had lived long in thoſe countries, whether they had ever heard that the inhabitants of Arabia and Ethiopia, &c. uſed locuſts as food. They anſwered that they had. He likewiſe aſked the ſame queſtion of Armenians, Cophtes, and Syrians, who lived in Arabia, and had travelled in Syria and near the Red-ſea; ſome of whom ſaid they heard of ſuch a practice, and others that they had often ſeen the people eat theſe inſects. He at laſt obtained complete ſatisfaction on this head from a learned Jbrek at Cairo, who had lived ſix years in Mocca. This gentleman beſides, in preſence of M. le Grand, the principal French interpreter at Cairo, and others, that a famine frequently rages at Mecca when there is a ſcarcity of corn in Egypt, which obliges the inhabitants to live upon coarſer food than ordinary: That when corn is ſcarce, the Arabians grind the locuſts in hand-mills, or ſtone mortars, and bake them into cakes, and uſe theſe cakes in place of bread: That he has frequently ſeen locuſts uſed by the Arabians, even when there was no ſcarcity of corn; but then they boil them, dreſs them with butter, and make them into a kind of fricaſſee, which he ſays is not diſagreeably taſted; for he had ſometimes taſted theſe locuſt-diſhes out of curioſity. From this account, we may fee the folly of that diſpute among divines about the nature of St John's food in the wilderneſs. Some of them ſay that locuſts were the fruits of certain trees, others that they were a kind of birds, &c.; but thoſe who adhered to the literal meaning of the text were at leaſt the moſt orthodox, although their arguments were perhaps not ſo ſtrong as they might have been, had they had an opportunity of quoting ſuch an author as Haſſelquiſt.

ACRI-

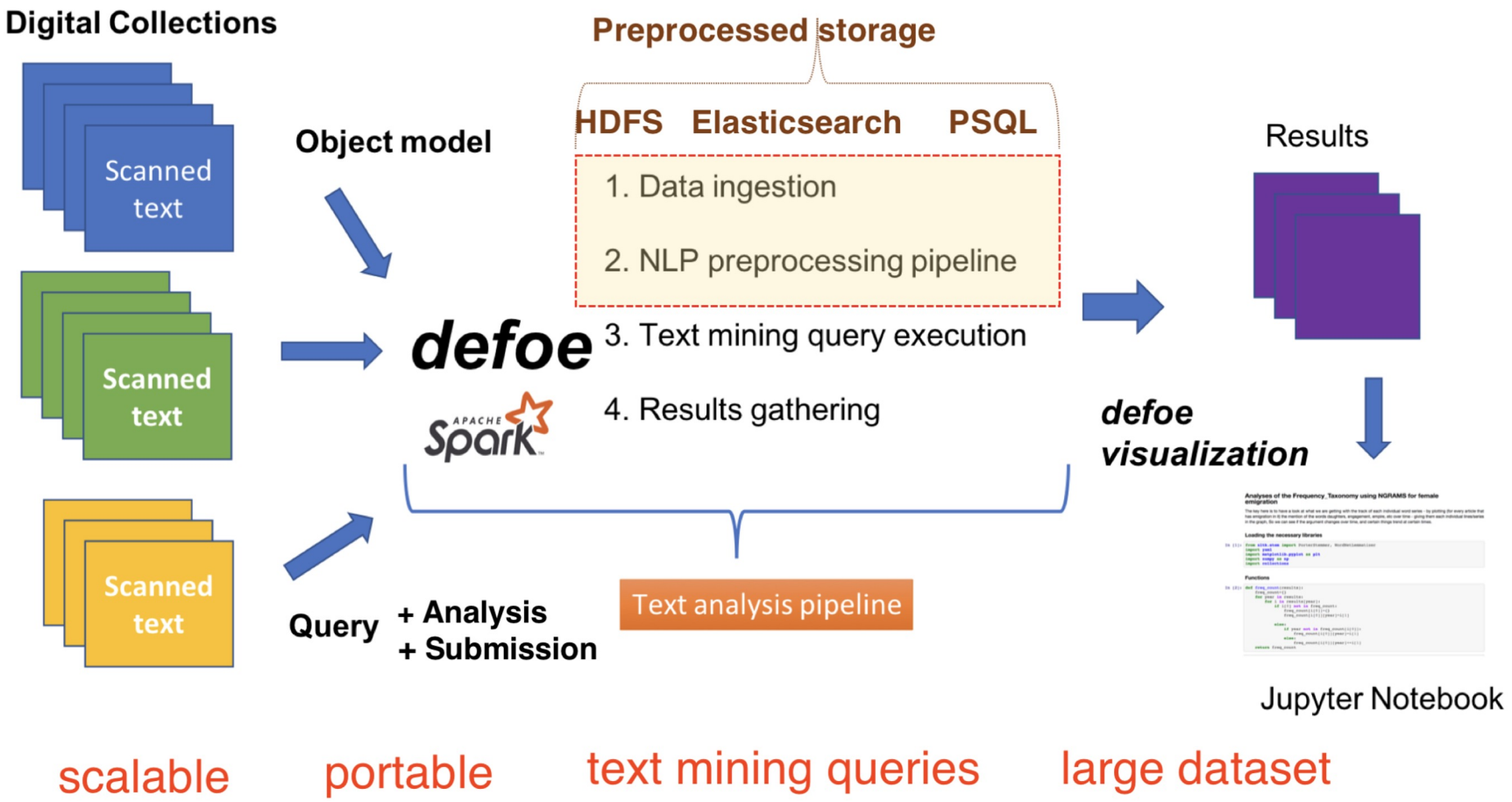
Page 36 of the First Edition of the Encyclopaedia Britannica

```
<?xml version="1.0" encoding="UTF-8"?>
<alto
    xmlns="http://www.loc.gov/standards/alto/v3/alto.xsd">
    <Description>
        <MeasurementUnit>pixel</MeasurementUnit>
        <sourceImageInformation>
            <fileName>./data/pdfs/c_188936619/i_144133901/188082693.23.pdf</fileName>
        </sourceImageInformation>
        <OCRProcessing ID="IdOcr">
            <ocrProcessingStep>
                <processingDateTime>Fri Jul 12 10:36:14 2019
</processingDateTime>
                <processingSoftware>
                    <softwareCreator>CONTRIBUTORS</softwareCreator>
                    <softwareName>pdfalto</softwareName>
                    <softwareVersion>0.1</softwareVersion>
                </processingSoftware>
            </ocrProcessingStep>
        </OCRProcessing>
    </Description>
    <Styles>
        <TextStyle ID="font0" FONTFAMILY="times" FONTSIZE="10.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE=""/>
        <TextStyle ID="font1" FONTFAMILY="times" FONTSIZE="6.500" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="subscript"/>
    </Styles>
    <Layout>
        <Page ID="Page36" PHYSICAL_IMG_NR="36" WIDTH="2466" HEIGHT="3327">
            <PrintSpace>
                <TextLine WIDTH="232" HEIGHT="75" ID="p36_t1" HPOS="512" VPOS="108">
                    <String ID="p36_w1" CONTENT="A" HPOS="512" VPOS="108" WIDTH="42" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="57" VPOS="108" HPOS="554"/>
                    <String ID="p36_w2" CONTENT="C" HPOS="611" VPOS="108" WIDTH="37" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="58" VPOS="108" HPOS="648"/>
                    <String ID="p36_w3" CONTENT="11" HPOS="706" VPOS="108" WIDTH="38" HEIGHT="75" STYLEREFS="font0"/>
                </TextLine>
                <TextLine WIDTH="132" HEIGHT="75" ID="p36_t2" HPOS="1119" VPOS="108">
                    <String ID="p36_w4" CONTENT="(" HPOS="1119" VPOS="108" WIDTH="19" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="62" VPOS="108" HPOS="1138"/>
                    <String ID="p36_w5" CONTENT="20" HPOS="1200" VPOS="108" WIDTH="51" HEIGHT="75" STYLEREFS="font0"/>
                </TextLine>
                <TextLine WIDTH="1029" HEIGHT="75" ID="p36_t3" HPOS="171" VPOS="182">
                    <String ID="p36_w6" CONTENT="ACQUEST," HPOS="171" VPOS="182" WIDTH="243" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="24" VPOS="182" HPOS="414"/>
                    <String ID="p36_w7" CONTENT="or" HPOS="438" VPOS="182" WIDTH="39" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="15" VPOS="182" HPOS="477"/>
                    <String ID="p36_w8" CONTENT="Acquist," HPOS="492" VPOS="182" WIDTH="193" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="24" VPOS="182" HPOS="685"/>
                    <String ID="p36_w9" CONTENT="inlaw," HPOS="709" VPOS="182" WIDTH="124" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="24" VPOS="182" HPOS="833"/>
                    <String ID="p36_w10" CONTENT="Gquifies" HPOS="857" VPOS="182" WIDTH="134" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="24" VPOS="182" HPOS="991"/>
                    <String ID="p36_w11" CONTENT="goods" HPOS="1015" VPOS="182" WIDTH="102" HEIGHT="75" STYLEREFS="font0"/>
                    <SP WIDTH="24" VPOS="182" HPOS="1117"/>
                    <String ID="p36_w12" CONTENT="got" HPOS="1141" VPOS="182" WIDTH="59" HEIGHT="75" STYLEREFS="font0"/>
                </TextLine>
                <TextLine WIDTH="792" HEIGHT="75" ID="p36_t4" HPOS="221" VPOS="223">
```

ALTO-XML of the Page 36 of the First Edition of the Encyclopaedia Britannica

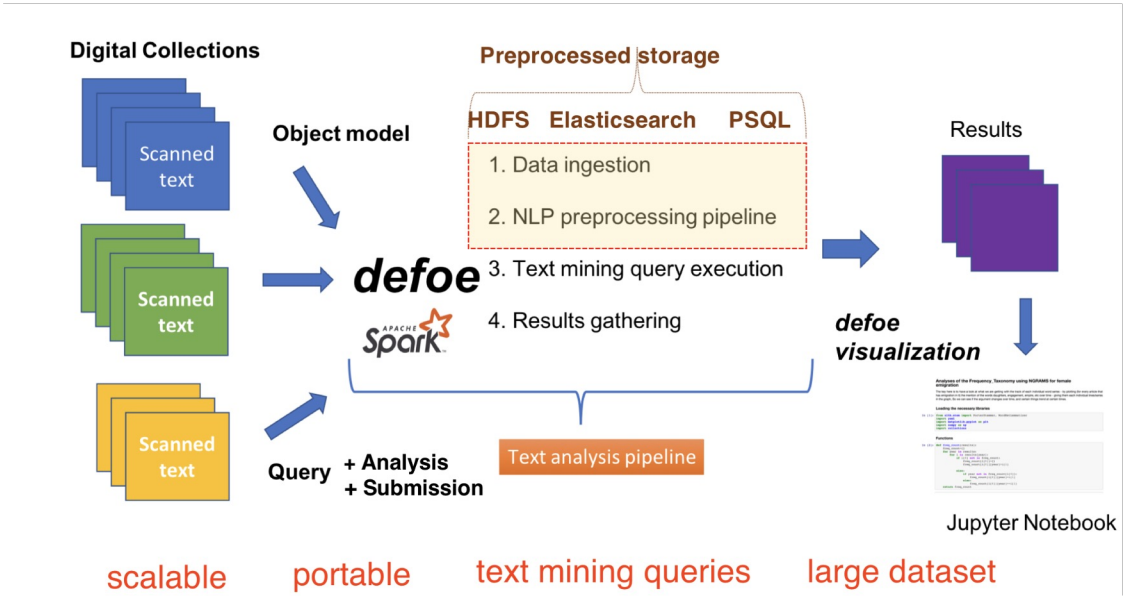
Phase 1.1: Information Extraction - EB

St Andrews / CS



https://github.com/francesNLP/defoe

Phase 1.1: Information Extraction - EB



Publications:

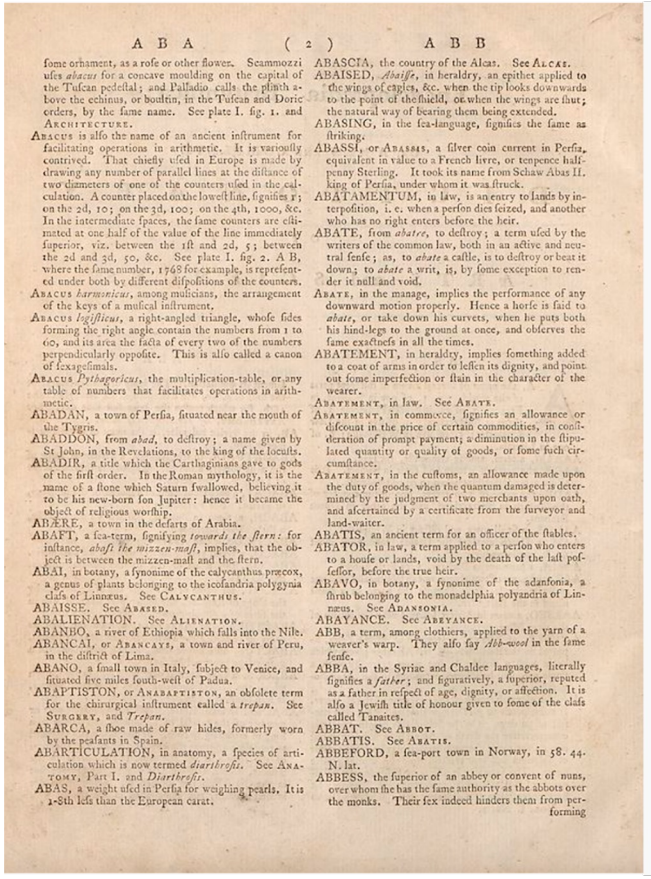
1. defoe: A Spark-based Toolbox for Analysing Digital Historical Textual Data, 2019
2. Geoparsing the historical gazetteers of scotland: Accurately computing location in mass digitised texts, 2020
3. Extending defoe for the efficient analysis of historical texts at scale, 2021

Phase 1.1: Information Extraction - EB

New *defoe* queries and heuristics

St Andrews / CS

- *defoe* query to **extract EB Terms:**
  - By page and classify them between **articles** and **topics** → Using ALTO XML



A B A (2) A B B

some ornament, as a rose or other flower. Scammozzi uses *abacus* for a concave moulding on the capital of the Tuscan pedestal; and Palladio calls the plinth above the echinus, or boultin, in the Tuscan and Doric orders, by the same name. See plate I. fig. 1. and ARCHITECTURE.

ABACUS is also the name of an ancient instrument for facilitating operations in arithmetic. It is variously contrived. That chiefly used in Europe is made by drawing any number of parallel lines at the distance of two diameters of one of the counters used in the calculation. A counter placed on the lowest line, signifies 1; on the 2d, 10; on the 3d, 100; on the 4th, 1000, &c. In the intermediate spaces, the same counters are estimated at one half of the value of the line immediately superior, viz. between the 1st and 2d, 5; between the 2d and 3d, 50, &c. See plate I. fig. 2. A B, where the same number, 1748 for example, is represented under both by different dispositions of the counters.

ABACUS *harmonicus*, among musicians, the arrangement of the keys of a musical instrument.

ABACUS *logisticus*, a right-angled triangle, whose sides forming the right angle contain the numbers from 1 to 60, and its area the facta of every two of the numbers perpendicularly opposite. This is also called a canon of sexagesimals.

ABACUS *Pythagoricus*, the multiplication-table, or any table of numbers that facilitates operations in arithmetic.

ABADAN, a town of Persia, situated near the mouth of the Tygris.

ABADDON, from *abad*, to destroy; a name given by St John, in the Revelations, to the king of the locusts.

ABADIR, a title which the Carthaginians gave to gods of the first order. In the Roman mythology, it is the name of a stone which Saturn swallowed, believing it to be his new-born son Jupiter: hence it became the object of religious worship.

ABÆRE, a town in the desarts of Arabia.

ABAFT, a sea-term, signifying *towards the stern*: for instance, *abaft the mizzen-mast*, implies, that the object is between the mizzen-mast and the stern.

ABAI, in botany, a synonime of the calycanthus præcox, a genus of plants belonging to the icosandria polygynia class of Linnæus. See CALYCANTHUS.

ABAISSE. See ABASED.

ABALIENATION. See ALIENATION.

ABANBO, a river of Ethiopia which falls into the Nile.

ABANCAI, or ABANCAYS, a town and river of Peru, in the district of Lima.

ABANO, a small town in Italy, subject to Venice, and situated five miles south-west of Padua.

ABAPTISTON, or ANABAPTISTON, an obsolete term for the chirurgical instrument called a *trepan*. See SURGERY, and *Trepan*.

ABARCA, a shoe made of raw hides, formerly worn by the peasants in Spain.

ABARTICULATION, in anatomy, a species of articulation which is now termed *diarthrosis*. See ANATOMY, Part I. and *Diarthrosis*.

ABAS, a weight used in Persia for weighing pearls. It is 1-8th less than the European carat.

ABASCIA, the country of the Alani. See ALANI.

ABAISED, *Abaissé*, in heraldry, an epithet applied to the wings of eagles, &c. when the tip looks downwards to the point of the shield, or when the wings are shut; the natural way of bearing them being extended.

ABASING, in the sea-language, signifies the same as striking.

ABASSI, or ABASSIS, a silver coin current in Persia, equivalent in value to a French livre, or tenpence halfpenny Sterling. It took its name from Schaw Abas II. king of Persia, under whom it was struck.

ABATAMENTUM, in law, is an entry to lands by interposition, i. e. when a person dies seized, and another who has no right enters before the heir.

ABATE, from *abatre*, to destroy; a term used by the writers of the common law, both in an active and neutral sense; as, to *abate* a castle, is to destroy or beat it down; to *abate* a writ, is, by some exception to render it null and void.

ABATE, in the manage, implies the performance of any downward motion properly. Hence a horse is said to *abate*, or take down his curvets, when he puts both his hind-legs to the ground at once, and observes the same exactness in all the times.

ABATEMENT, in heraldry, implies something added to a coat of arms in order to lessen its dignity, and point out some imperfection or stain in the character of the wearer.

ABATEMENT, in law. See ABATE.

ABATEMENT, in commerce, signifies an allowance or discount in the price of certain commodities, in consideration of prompt payment; a diminution in the stipulated quantity or quality of goods, or some such circumstance.

ABATEMENT, in the customs, an allowance made upon the duty of goods, when the quantum damaged is determined by the judgment of two merchants upon oath, and ascertained by a certificate from the surveyor and land-waiter.

ABATIS, an ancient term for an officer of the stables.

ABATOR, in law, a term applied to a person who enters to a house or lands, void by the death of the last possessor, before the true heir.

ABAVO, in botany, a synonime of the adansonia, a shrub belonging to the monadelphia polyandria of Linnæus. See ADANSONIA.

ABAYANCE. See ABEYANCE.

ABB, a term, among clothiers, applied to the yarn of a weaver's warp. They also say *Abb-wood* in the same sense.

ABBA, in the Syriac and Chaldee languages, literally signifies a *father*; and figuratively, a superior, reputed as a father in respect of age, dignity, or affection. It is also a Jewish title of honour given to some of the class called Tanaites.

ABBAT. See ABBOT.

ABBATIS. See ABATIS.

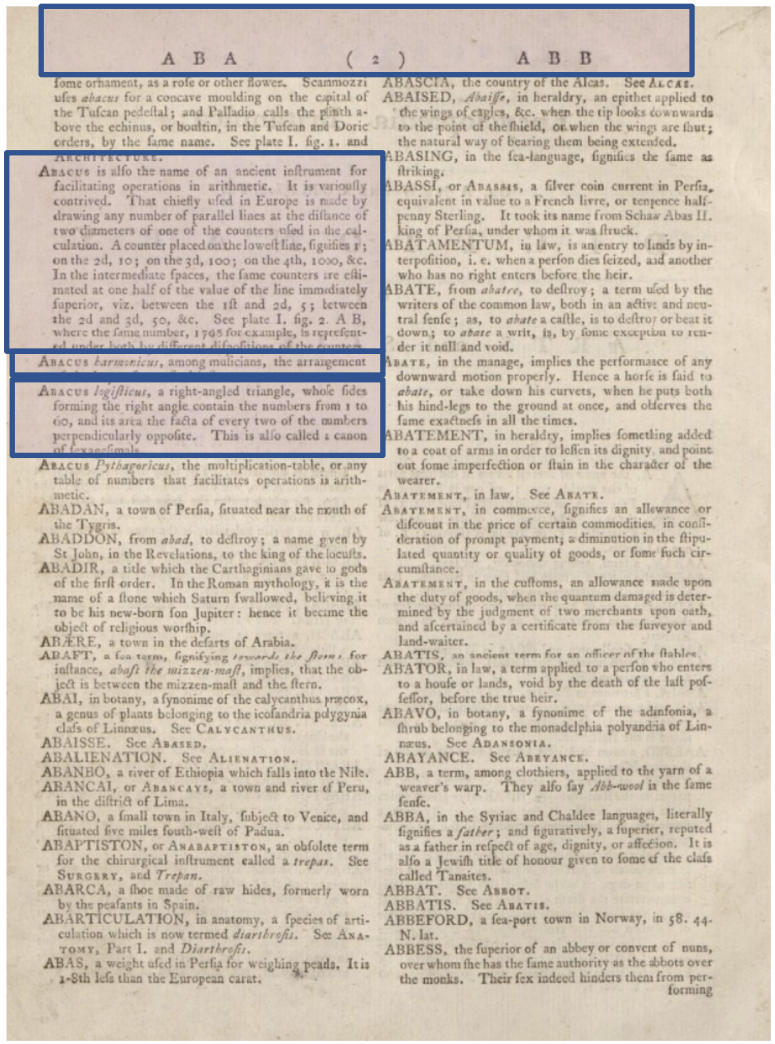
ABBEFORD, a sea-port town in Norway, in 58. 44. N. lat.

ABBESS, the superior of an abbey or convent of nuns, over whom she has the same authority as the abbots over the monks. Their sex indeed hinders them from performing

Phase 1.1: Information Extraction - EB

New *defoe* queries and heuristics

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 + ","



A B A (2) A B B

some ornament, as a rose or other flower. Scammozzi uses *abacus* for a concave moulding on the capital of the Tuscan pedestal; and Palladio calls the plinth above the echinus, or bolstin, in the Tuscan and Doric orders, by the same name. See plate I. fig. 1. and ARCHITECTURE.

ABACUS is also the name of an ancient instrument for facilitating operations in arithmetic. It is variously contrived. That chiefly used in Europe is made by drawing any number of parallel lines at the distance of two diameters of one of the counters used in the calculation. A counter placed on the lowest line, signifies 1; on the 2d, 10; on the 3d, 100; on the 4th, 1000, &c. In the intermediate spaces, the same counters are estimated at one half of the value of the line immediately superior, viz. between the 1st and 2d, 5; between the 2d and 3d, 50, &c. See plate I. fig. 2. A B, where the same number, 1765 for example, is represented under both the different dispositions of the counters.

ABACUS *harmonicus*, among musicians, the arrangement

ABACUS *logisticus*, a right-angled triangle, whose sides forming the right angle contain the numbers from 1 to 60, and its area the facta of every two of the numbers perpendicularly opposite. This is also called a canon of sexagesimals.

ABACUS *Pythagoricus*, the multiplication-table, or any table of numbers that facilitates operations in arithmetic.

ABADAN, a town of Persia, situated near the mouth of the Tygris.

ABADDON, from *abad*, to destroy; a name given by St John, in the Revelations, to the king of the locusts.

ABADIR, a title which the Carthaginians gave to gods of the first order. In the Roman mythology, it is the name of a stone which Saturn swallowed, believing it to be his new-born son Jupiter: hence it became the object of religious worship.

ABÆRE, a town in the desarts of Arabia.

ABAFT, a sea term, signifying *towards the stern*: for instance, *abaft the mizzen-mast*, implies, that the object is between the mizzen-mast and the stern.

ABAI, in botany, a synonime of the calycanthus præcox, a genus of plants belonging to the icosandria polygynia class of Linnæus. See CALYCANTHUS.

ABAISSE. See ABASED.

ABALIENATION. See ALIENATION.

ABANBO, a river of Ethiopia which falls into the Nile.

ABANCAI, or ABANCAYS, a town and river of Peru, in the district of Lima.

ABANO, a small town in Italy, subject to Venice, and situated five miles south-west of Padua.

ABAPTISTON, or ANABAPTISTON, an obsolete term for the chirurgical instrument called a *trepan*. See SURGERY, and *Trepan*.

ABARCA, a shoe made of raw hides, formerly worn by the peasants in Spain.

ABARTICULATION, in anatomy, a species of articulation which is now termed *diarthrosis*. See ANATOMY, Part I. and *Diarthrosis*.

ABAS, a weight used in Persia for weighing pearls. It is 1-8th less than the European carat.

ABASCIA, the country of the Alans. See ALANS.

ABAISED, *Abaissé*, in heraldry, an epithet applied to the wings of eagles, &c. when the tip looks downwards to the point of the shield, or when the wings are shut; the natural way of bearing them being extended.

ABASING, in the sea-language, signifies the same as striking.

ABASSI, or ABASSIS, a silver coin current in Persia, equivalent in value to a French livre, or tenpence halfpenny Sterling. It took its name from Schaw Abas II. king of Persia, under whom it was struck.

ABATAMENTUM, in law, is an entry to lands by interposition, i. e. when a person dies seized, and another who has no right enters before the heir.

ABATE, from *abatre*, to destroy; a term used by the writers of the common law, both in an active and neutral sense; as, to *abate* a castle, is to destroy or beat it down; to *abate* a writ, is, by some exception to render it null and void.

ABATE, in the manage, implies the performance of any downward motion properly. Hence a horse is said to *abate*, or take down his curvets, when he puts both his hind-legs to the ground at once, and observes the same exactness in all the times.

ABATEMENT, in heraldry, implies something added to a coat of arms in order to lessen its dignity, and point out some imperfection or stain in the character of the wearer.

ABATEMENT, in law. See ABATE.

ABATEMENT, in commerce, signifies an allowance or discount in the price of certain commodities, in consideration of prompt payment; a diminution in the stipulated quantity or quality of goods, or some such circumstance.

ABATEMENT, in the customs, an allowance made upon the duty of goods, when the quantum damaged is determined by the judgment of two merchants upon oath, and ascertained by a certificate from the surveyor and land-waiter.

ABATIS, an ancient term for an officer of the stables.

ABATOR, in law, a term applied to a person who enters to a house or lands, void by the death of the last possessor, before the true heir.

ABAVO, in botany, a synonime of the adansonia, a shrub belonging to the monadelphia polyandria of Linnæus. See ADANSONIA.

ABAYANCE. See ABEYANCE.

ABB, a term, among clothiers, applied to the yarn of a weaver's warp. They also say *Abb-wool* in the same sense.

ABBA, in the Syriac and Chaldee languages, literally signifies a *father*; and figuratively, a superior, reputed as a father in respect of age, dignity, or affection. It is also a Jewish title of honour given to some of the class called Tanaites.

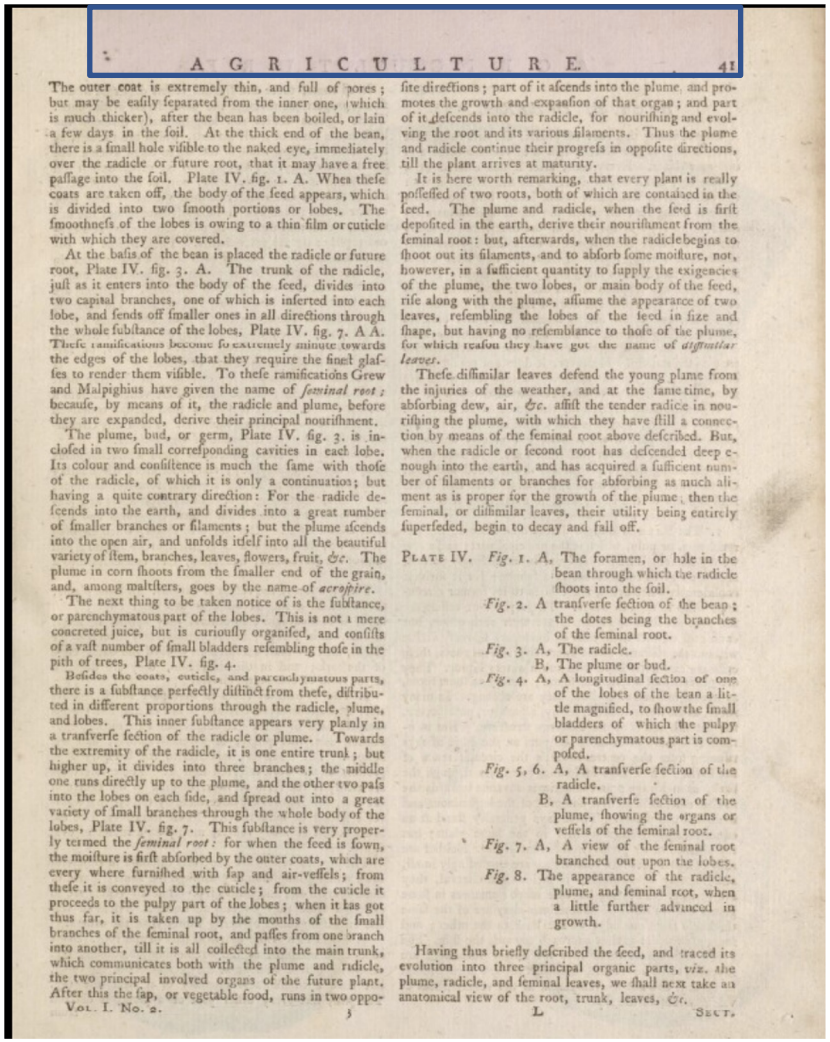
ABBAT. See ABBOT.

ABBATIS. See ABATIS.

ABBEFORD, a sea-port town in Norway, in 58. 44. N. lat.

ABBESS, the superior of an abbey or convent of nuns, over whom she has the same authority as the abbots over the monks. Their sex indeed hinders them from performing

Articles



A G R I C U L T U R E. 41

The outer coat is extremely thin, and full of pores; but may be easily separated from the inner one, (which is much thicker), after the bean has been boiled, or lain a few days in the soil. At the thick end of the bean, there is a small hole visible to the naked eye, immediately over the radicle or future root, that it may have a free passage into the soil. Plate IV. fig. 1. A. When these coats are taken off, the body of the seed appears, which is divided into two smooth portions or lobes. The smoothness of the lobes is owing to a thin film or cuticle with which they are covered.

At the basis of the bean is placed the radicle or future root, Plate IV. fig. 3. A. The trunk of the radicle, just as it enters into the body of the seed, divides into two capital branches, one of which is inserted into each lobe, and sends off smaller ones in all directions through the whole substance of the lobes, Plate IV. fig. 7. A A. These ramifications become so extremely minute towards the edges of the lobes, that they require the finest glasses to render them visible. To these ramifications Grew and Malpighius have given the name of *seminal root*; because, by means of it, the radicle and plume, before they are expanded, derive their principal nourishment.

The plume, bud, or germ, Plate IV. fig. 3. is inclosed in two small corresponding cavities in each lobe. Its colour and consistence is much the same with those of the radicle, of which it is only a continuation; but having a quite contrary direction: For the radicle descends into the earth, and divides into a great number of smaller branches or filaments; but the plume ascends into the open air, and unfolds itself into all the beautiful variety of stem, branches, leaves, flowers, fruit, &c. The plume in corn shoots from the smaller end of the grain, and, among maltsters, goes by the name of *acrospire*.

The next thing to be taken notice of is the substance, or parenchymatous part of the lobes. This is not a mere concreted juice, but is curiously organised, and consists of a vast number of small bladders resembling those in the pith of trees, Plate IV. fig. 4.

Besides the coats, cuticle, and parenchymatous parts, there is a substance perfectly distinct from these, distributed in different proportions through the radicle, plume, and lobes. This inner substance appears very plainly in a transverse section of the radicle or plume. Towards the extremity of the radicle, it is one entire trunk; but higher up, it divides into three branches; the middle one runs directly up to the plume, and the other two pass into the lobes on each side, and spread out into a great variety of small branches through the whole body of the lobes, Plate IV. fig. 7. This substance is very properly termed the *seminal root*: for when the seed is sown, the moisture is first absorbed by the outer coats, which are every where furnished with sap and air-vessels; from these it is conveyed to the cuticle; from the cuticle it proceeds to the pulpy part of the lobes; when it has got thus far, it is taken up by the mouths of the small branches of the seminal root, and passes from one branch into another, till it is all collected into the main trunk, which communicates both with the plume and radicle, the two principal involved organs of the future plant. After this the sap, or vegetable food, runs in two opposite directions; part of it ascends into the plume, and promotes the growth and expansion of that organ; and part of it descends into the radicle, for nourishing and evolving the root and its various filaments. Thus the plume and radicle continue their progress in opposite directions, till the plant arrives at maturity.

It is here worth remarking, that every plant is really possessed of two roots, both of which are contained in the seed. The plume and radicle, when the seed is first deposited in the earth, derive their nourishment from the seminal root: but, afterwards, when the radicle begins to shoot out its filaments, and to absorb some moisture, not, however, in a sufficient quantity to supply the exigencies of the plume, the two lobes, or main body of the seed, rise along with the plume, assume the appearance of two leaves, resembling the lobes of the seed in size and shape, but having no resemblance to those of the plume, for which reason they have got the name of *stipulular leaves*.

These dissimilar leaves defend the young plume from the injuries of the weather, and at the same time, by absorbing dew, air, &c. assist the tender radicle in nourishing the plume, with which they have still a connection by means of the seminal root above described. But, when the radicle or second root has descended deep enough into the earth, and has acquired a sufficient number of filaments or branches for absorbing as much aliment as is proper for the growth of the plume, then the seminal, or dissimilar leaves, their utility being entirely superseded, begin to decay and fall off.

PLATE IV. *Fig.* 1. A, The foramen, or hole in the bean through which the radicle shoots into the soil.
Fig. 2. A transverse section of the bean; the dotes being the branches of the seminal root.
Fig. 3. A, The radicle.
B, The plume or bud.
Fig. 4. A, A longitudinal section of one of the lobes of the bean a little magnified, to show the small bladders of which the pulpy or parenchymatous part is composed.
Fig. 5, 6. A, A transverse section of the radicle.
B, A transverse section of the plume, showing the organs or vessels of the seminal root.
Fig. 7. A, A view of the seminal root branched out upon the lobes.
Fig. 8. The appearance of the radicle, plume, and seminal root, when a little further advanced in growth.

Having thus briefly described the seed, and traced its evolution into three principal organic parts, viz. the plume, radicle, and seminal leaves, we shall next take an anatomical view of the root, trunk, leaves, &c.

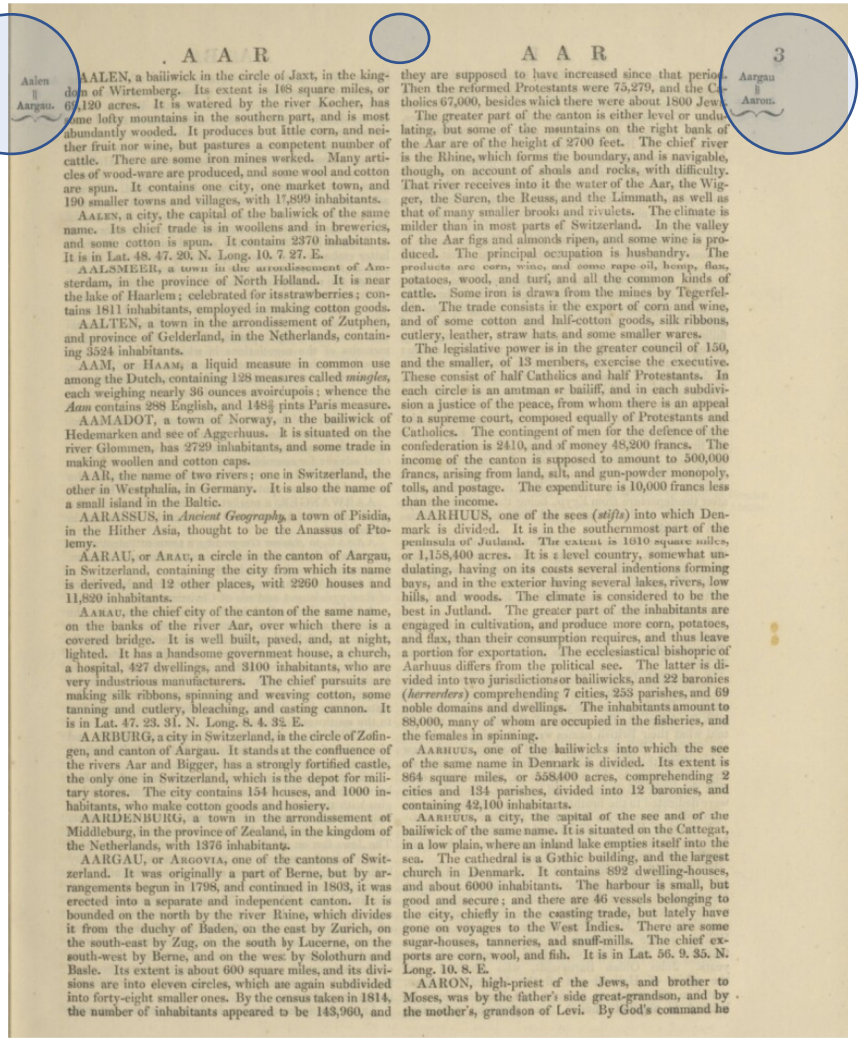
Topic

Edition 1 - 1771

Phase 1.1: Information Extraction - EB

New *defoe* queries and heuristics

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 + “,”



. A A R

Aalen ‖ Aargau.

AALEN, a bailiwick in the circle of Jaxt, in the kingdom of Wirtemberg. Its extent is 168 square miles, or 67,120 acres. It is watered by the river Kocher, has some lofty mountains in the southern part, and is most abundantly wooded. It produces but little corn, and neither fruit nor wine, but pastures a competent number of cattle. There are some iron mines worked. Many articles of wood-ware are produced, and some wool and cotton are spun. It contains one city, one market town, and 190 smaller towns and villages, with 17,899 inhabitants.

AALEN, a city, the capital of the baliwick of the same name. Its chief trade is in woollens and in breweries, and some cotton is spun. It contains 2370 inhabitants. It is in Lat. 48. 47. 20. N. Long. 10. 7. 27. E.

AALSMEER, a town in the arrodissement of Amsterdam, in the province of North Holland. It is near the lake of Haarlem; celebrated for itsstrawberries; contains 1811 inhabitants, employed in making cotton goods.

AALTEN, a town in the arrondissement of Zutphen, and province of Gelderland, in the Netherlands, containing 3524 inhabitants.

AAM, or HAAM, a liquid measure in common use among the Dutch, containing 128 measures called *mingles*, each weighing nearly 36 ounces avoirdupois; whence the *Aam* contains 298 English, and 148⅝ pints Paris measure.

AAMADOT, a town of Norway, in the bailiwick of Hedemarken and see of Aggerhuus. It is situated on the river Glommen, has 2729 inhabitants, and some trade in making woollen and cotton caps.

AAR, the name of two rivers; one in Switzerland, the other in Westphalia, in Germany. It is also the name of a small island in the Baltic.

AARASSUS, in *Ancient Geography*, a town of Pisidia, in the Hither Asia, thought to be the Anassus of Ptolemy.

AARAU, or ARAU, a circle in the canton of Aargau, in Switzerland, containing the city from which its name is derived, and 12 other places, with 2260 houses and 11,820 inhabitants.

AARAU, the chief city of the canton of the same name, on the banks of the river Aar, over which there is a covered bridge. It is well built, paved, and, at night, lighted. It has a handsome government house, a church, a hospital, 427 dwellings, and 3100 inhabitants, who are very industrious manufacturers. The chief pursuits are making silk ribbons, spinning and weaving cotton, some tanning and cutlery, bleaching, and casting cannon. It is in Lat. 47. 23. 31. N. Long. 8. 4. 32. E.

AARBURG, a city in Switzerland, in the circle of Zofingen, and canton of Aargau. It stands at the confluence of the rivers Aar and Bigger, has a strongly fortified castle, the only one in Switzerland, which is the depot for military stores. The city contains 154 houses, and 1000 inhabitants, who make cotton goods and hosiery.

AARDENBURG, a town in the arrondissement of Middleburg, in the province of Zealand, in the kingdom of the Netherlands, with 1376 inhabitants.

AARGAU, or ARGOVIA, one of the cantons of Switzerland. It was originally a part of Berne, but by arrangements begun in 1798, and continued in 1803, it was erected into a separate and indepencent canton. It is bounded on the north by the river Raine, which divides it from the duchy of Baden, on the east by Zurich, on the south-east by Zug, on the south by Lucerne, on the south-west by Berne, and on the west by Solothurn and Basle. Its extent is about 600 square miles, and its divisions are into eleven circles, which are again subdivided into forty-eight smaller ones. By the census taken in 1814, the number of inhabitants appeared to be 143,960, and

A A R 3

Aargau ‖ Aaron.

they are supposed to have increased since that period. Then the reformed Protestants were 75,279, and the Catholics 67,000, besides which there were about 1800 Jews.

The greater part of the canton is either level or undulating, but some of the mountains on the right bank of the Aar are of the height of 2700 feet. The chief river is the Rhine, which forms the boundary, and is navigable, though, on account of shoals and rocks, with difficulty. That river receives into it the water of the Aar, the Wigger, the Suren, the Reuss, and the Limmath, as well as that of many smaller brooks and rivulets. The climate is milder than in most parts of Switzerland. In the valley of the Aar figs and almonds ripen, and some wine is produced. The principal occupation is husbandry. The products are corn, wine, and some rape oil, hemp, flax, potatoes, wood, and turf, and all the common kinds of cattle. Some iron is drawn from the mines by Tegerfelden. The trade consists in the export of corn and wine, and of some cotton and half-cotton goods, silk ribbons, cutlery, leather, straw hats and some smaller wares.

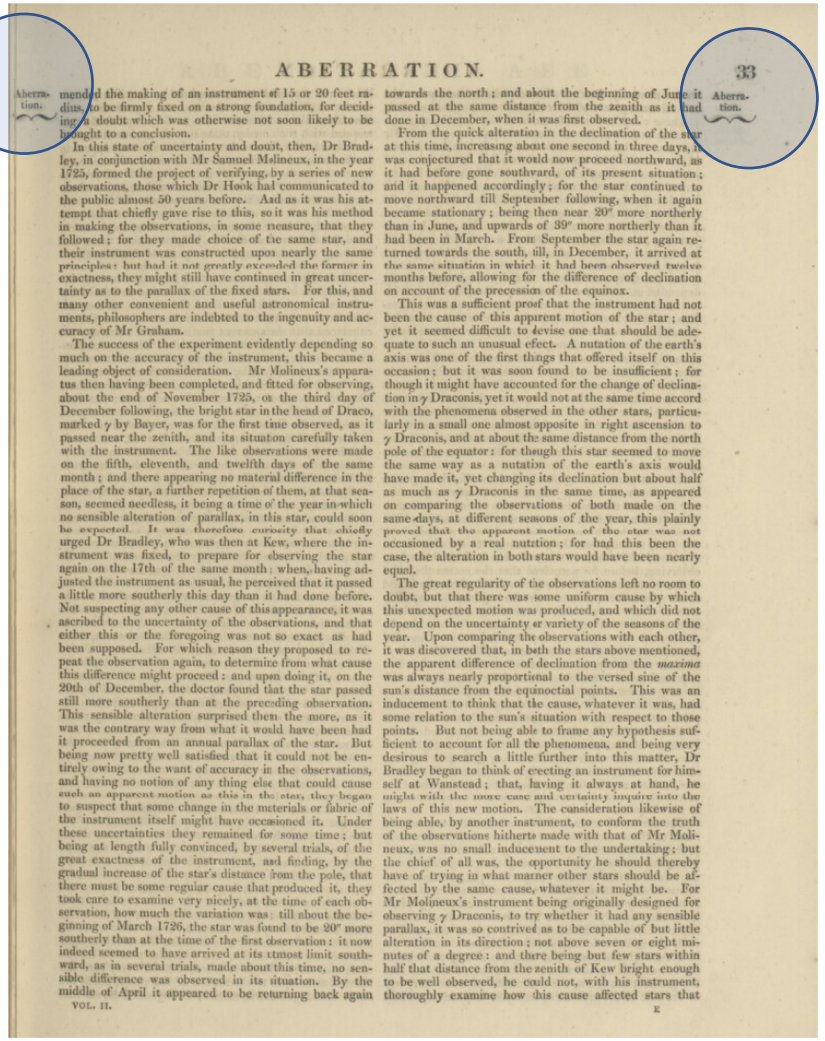
The legislative power is in the greater council of 150, and the smaller, of 13 members, exercise the executive. These consist of half Catholics and half Protestants. In each circle is an amtman or bailiff, and in each subdivision a justice of the peace, from whom there is an appeal to a supreme court, composed equally of Protestants and Catholics. The contingent of men for the defence of the confederation is 2410, and of money 48,200 francs. The income of the canton is supposed to amount to 500,000 francs, arising from land, salt, and gun-powder monopoly, tolls, and postage. The expenditure is 10,000 francs less than the income.

AARHUUS, one of the sees (*stifts*) into which Denmark is divided. It is in the southernmost part of the peninsula of Jutland. The extent is 1610 square miles, or 1,158,400 acres. It is a level country, somewhat undulating, having on its coasts several indentions forming bays, and in the exterior having several lakes, rivers, low hills, and woods. The climate is considered to be the best in Jutland. The greater part of the inhabitants are engaged in cultivation, and produce more corn, potatoes, and flax, than their consumption requires, and thus leave a portion for exportation. The ecclesiastical bishopric of Aarhuus differs from the political see. The latter is divided into two jurisdictions or bailiwicks, and 22 baronies (*herrerders*) comprehending 7 cities, 253 parishes, and 69 noble domains and dwellings. The inhabitants amount to 88,000, many of whom are occupied in the fisheries, and the females in spinning.

AARHUUS, one of the bailiwicks into which the see of the same name in Denmark is divided. Its extent is 864 square miles, or 558,400 acres, comprehending 2 cities and 134 parishes, divided into 12 baronies, and containing 42,100 inhabitants.

AARHUUS, a city, the capital of the see and of the bailiwick of the same name. It is situated on the Cattegat, in a low plain, where an inland lake empties itself into the sea. The cathedral is a Gothic building, and the largest church in Denmark. It contains 892 dwelling-houses, and about 6000 inhabitants. The harbour is small, but good and secure; and there are 46 vessels belonging to the city, chiefly in the coasting trade, but lately have gone on voyages to the West Indies. There are some sugar-houses, tanneries, and snuff-mills. The chief exports are corn, wool, and fish. It is in Lat. 56. 9. 35. N. Long. 10. 8. E.

AARON, high-priest of the Jews, and brother to Moses, was by the father's side great-grandson, and by the mother's, grandson of Levi. By God's command he



ABERRATION. 33

Aberration. Aberration.

mended the making of an instrument of 15 or 20 feet radius, to be firmly fixed on a strong foundation, for deciding a doubt which was otherwise not soon likely to be brought to a conclusion.

In this state of uncertainty and doubt, then, Dr Bradley, in conjunction with Mr Samuel Molineux, in the year 1725, formed the project of verifying, by a series of new observations, those which Dr Hook had communicated to the public almost 50 years before. And as it was his attempt that chiefly gave rise to this, so it was his method in making the observations, in some measure, that they followed; for they made choice of the same star, and their instrument was constructed upon nearly the same principles: but had it not greatly exceeded the former in exactness, they might still have continued in great uncertainty as to the parallax of the fixed stars. For this, and many other convenient and useful astronomical instruments, philosophers are indebted to the ingenuity and accuracy of Mr Graham.

The success of the experiment evidently depending so much on the accuracy of the instrument, this became a leading object of consideration. Mr Molineux's apparatus then having been completed, and fitted for observing, about the end of November 1725, on the third day of December following, the bright star in the head of Draco, marked γ by Bayer, was for the first time observed, as it passed near the zenith, and its situation carefully taken with the instrument. The like observations were made on the fifth, eleventh, and twelfth days of the same month; and there appearing no material difference in the place of the star, a further repetition of them, at that season, seemed needless, it being a time of the year in which no sensible alteration of parallax, in this star, could soon be expected. It was therefore curiosity that chiefly urged Dr Bradley, who was then at Kew, where the instrument was fixed, to prepare for observing the star again on the 17th of the same month: when, having adjusted the instrument as usual, he perceived that it passed a little more southerly this day than it had done before. Not suspecting any other cause of this appearance, it was ascribed to the uncertainty of the observations, and that either this or the foregoing was not so exact as had been supposed. For which reason they proposed to repeat the observation again, to determine from what cause this difference might proceed: and upon doing it, on the 20th of December, the doctor found that the star passed still more southerly than at the preceding observation. This sensible alteration surprised them the more, as it was the contrary way from what it would have been had it proceeded from an annual parallax of the star. But being now pretty well satisfied that it could not be entirely owing to the want of accuracy in the observations, and having no notion of any thing else that could cause such an apparent motion as this in the star, they began to suspect that some change in the materials or fabric of the instrument itself might have occasioned it. Under these uncertainties they remained for some time; but being at length fully convinced, by several trials, of the great exactness of the instrument, and finding, by the gradual increase of the star's distance from the pole, that there must be some regular cause that produced it, they took care to examine very nicely, at the time of each observation, how much the variation was: till about the beginning of March 1726, the star was found to be 20″ more southerly than at the time of the first observation: it now indeed seemed to have arrived at its utmost limit southward, as in several trials, made about this time, no sensible difference was observed in its situation. By the middle of April it appeared to be returning back again towards the north; and about the beginning of June it passed at the same distance from the zenith as it had done in December, when it was first observed.

From the quick alteration in the declination of the star at this time, increasing about one second in three days, it was conjectured that it would now proceed northward, as it had before gone southward, of its present situation; and it happened accordingly; for the star continued to move northward till September following, when it again became stationary; being then near 20″ more northerly than in June, and upwards of 39″ more northerly than it had been in March. From September the star again returned towards the south, till, in December, it arrived at the same situation in which it had been observed twelve months before, allowing for the difference of declination on account of the precession of the equinox.

This was a sufficient proof that the instrument had not been the cause of this apparent motion of the star; and yet it seemed difficult to devise one that should be adequate to such an unusual effect. A nutation of the earth's axis was one of the first things that offered itself on this occasion; but it was soon found to be insufficient; for though it might have accounted for the change of declination in γ Draconis, yet it would not at the same time accord with the phenomena observed in the other stars, particularly in a small one almost opposite in right ascension to γ Draconis, and at about the same distance from the north pole of the equator: for though this star seemed to move the same way as a nutation of the earth's axis would have made it, yet changing its declination but about half as much as γ Draconis in the same time, as appeared on comparing the observations of both made on the same days, at different seasons of the year, this plainly proved that the apparent motion of the star was not occasioned by a real nutation; for had this been the case, the alteration in both stars would have been nearly equal.

The great regularity of the observations left no room to doubt, but that there was some uniform cause by which this unexpected motion was produced, and which did not depend on the uncertainty or variety of the seasons of the year. Upon comparing the observations with each other, it was discovered that, in both the stars above mentioned, the apparent difference of declination from the *maxima* was always nearly proportional to the versed sine of the sun's distance from the equinoctial points. This was an inducement to think that the cause, whatever it was, had some relation to the sun's situation with respect to those points. But not being able to frame any hypothesis sufficient to account for all the phenomena, and being very desirous to search a little further into this matter, Dr Bradley began to think of erecting an instrument for himself at Wanstead; that, having it always at hand, he might with the more ease and certainty inquire into the laws of this new motion. The consideration likewise of being able, by another instrument, to conform the truth of the observations hitherto made with that of Mr Molineux, was no small inducement to the undertaking; but the chief of all was, the opportunity he should thereby have of trying in what manner other stars should be affected by the same cause, whatever it might be. For Mr Molineux's instrument being originally designed for observing γ Draconis, to try whether it had any sensible parallax, it was so contrived as to be capable of but little alteration in its direction; not above seven or eight minutes of a degree: and there being but few stars within half that distance from the zenith of Kew bright enough to be well observed, he could not, with his instrument, thoroughly examine how this cause affected stars that

Articles

Topic

Edition 7

New *defoe* queries and heuristics


```
term                                                          SCIENCE
definition          in philolophy, denotes any dpdfrine, deduced f...
relatedTerms                                                      []
header                                                        SCISCO
startsAt                                                         658
endsAt                                                           658
numberOfTerms                                                     24
numberOfWords                                                     15
numberOfPages                                                    872
positionPage                                                       7
typeTerm                                                     Article
editionTitle                        First edition, 1771, Volume 3, M-Z
editionNum                                                         1
supplementTitle
supplementsTo                                                     []
year                                                            1771
place                                                       Edinburgh
volumeTitle         Encyclopaedia Britannica; or, A dictionary of ...
volumeNum                                                          3
letters                                                          M-Z
part                                                               0
altoXML                                 144133903/alto/144812443.34.xml
Name: 7454, dtype: object
```

Science term information extracted from the First Edition - 1771

Phase 1.1: Information Extraction - EB

New *defoe* queries and heuristics

St Andrews / CS

- *defoe* query to extract EB **metadata :**
  - Metadata per Edition and Volume → Using METS XML

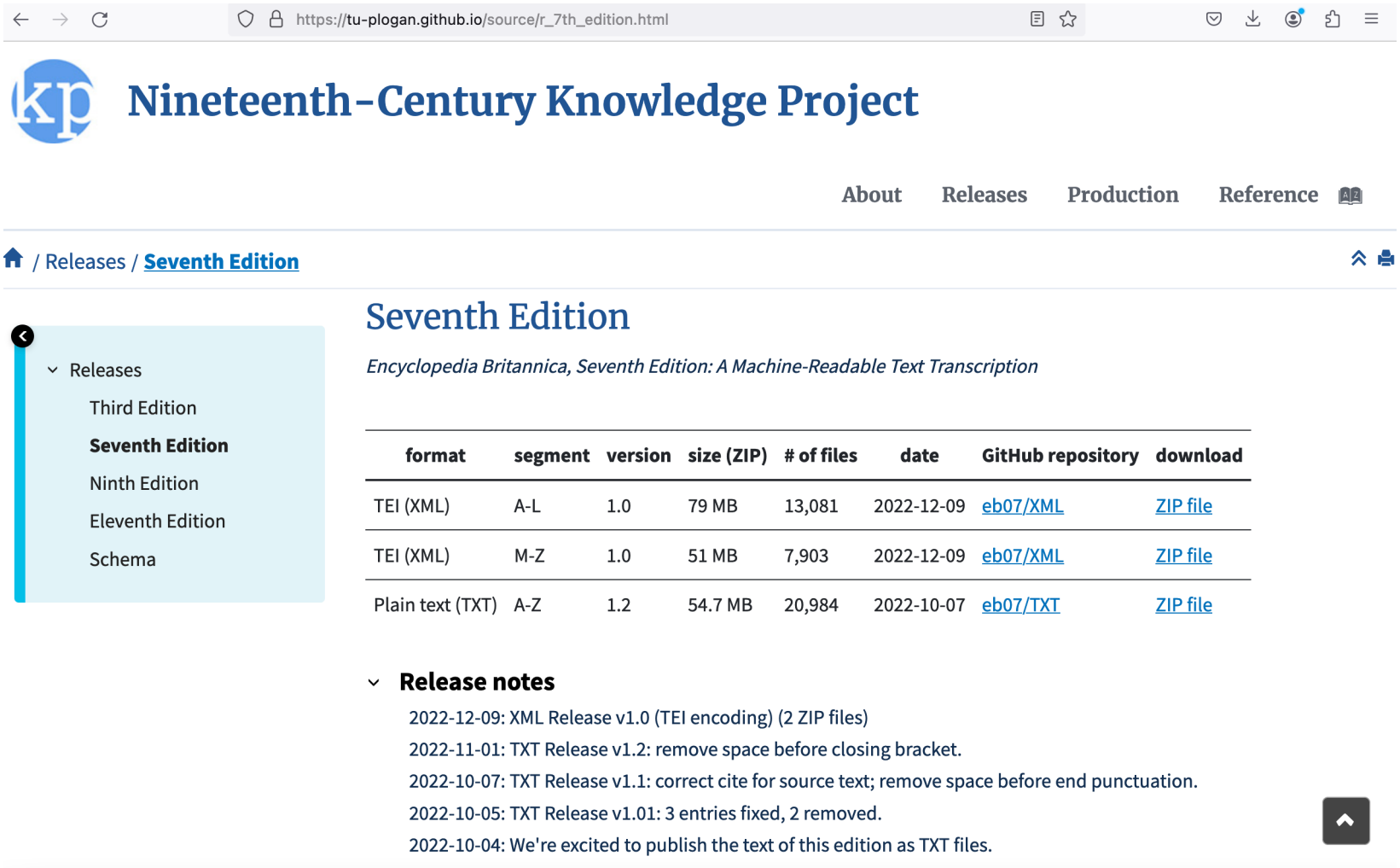
	MMSID	editionTitle	editor	editor_date	genre	language	termsOfAddress	numberOfPages	physicalDescription	place	...	
0	992277653804341	First edition, 1771, Volume 1, A-B	Smellie, William	1740-1795	encyclopedia	eng	None	832	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	...	h
1	992277653804341	First edition, 1771, Volume 2, C-L	Smellie, William	1740-1795	encyclopedia	eng	None	1018	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	...	h
2	992277653804341	First edition, 1771, Volume 3, M-Z	Smellie, William	1740-1795	encyclopedia	eng	None	872	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	...	h
3	9929192893804340	First edition, 1773, Volume 1, A-B	Smellie, William	1740-1795	encyclopedia	eng	None	844	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	...	h
4	9929192893804340	First edition, 1773, Volume 2, C-L	Smellie, William	1740-1795	encyclopedia	eng	None	1032	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	...	h
5	9929192893804340	First edition, 1773, Volume 3, M-Z	Smellie, William	1740-1795	encyclopedia	eng	None	864	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	...	h

Subset of **metadata extracted** for the volumes of the **First Edition**. This edition is a 3-volume reference work, **issued twice, in 1771 and 1773.**

Phase 1.1: Information Extraction - EB

Work in progress- Using the clean data from the 7th Edition - KP

St Andrews / CS



Phase 1.2: Information Extraction – Other NLS Collections

St Andrews / CS

- **Chapbooks printed in Scotland:** This dataset includes over **3,000 chapbooks** printed in Scotland. Popular reading materials from the late **17th to the 19th century**. They were usually printed on a single sheet and then **folded into books of 8, 12, 16 and 24 pages**. **47,329 ALTO XMLs** at page level,~10 million OCRed words.

- **Ladies' Edinburgh Debating Society** : This collection comprises two Edinburgh journals: *The Attempt and The Ladies Edinburgh* Magazine. It spans **16 volumes** from **19th century**. **6,354 ALTO XMLs** at the page level ~2.5 million words.

- **Gazetteers of Scotland collection**: **20 volumes** historical gazetteers of Scotland from the **19th century**. **~13,000 ALTO-XMLs** at page level, ~14.5 million words.

New *defoe* heuristic approach – write_metadata_pages.yml :

- parallel extraction of text from each page using ALTO-XML files
- extracting metadata from both the collection and volume using METS-XML files

Structuring the extracted information.

- For each volume - permanent URL that allows visualization of the page images.

Phase 1.2: Information Extraction – Other NLS Collections

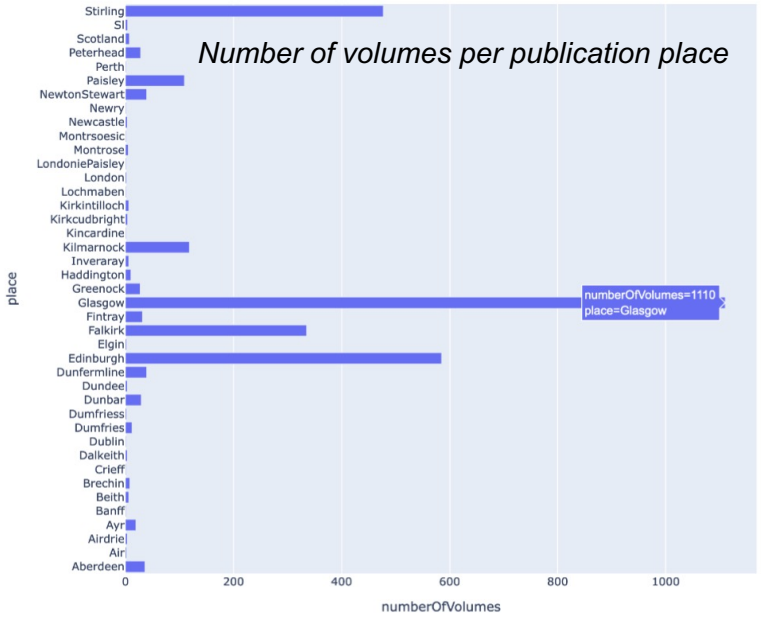
St Andrews / CS

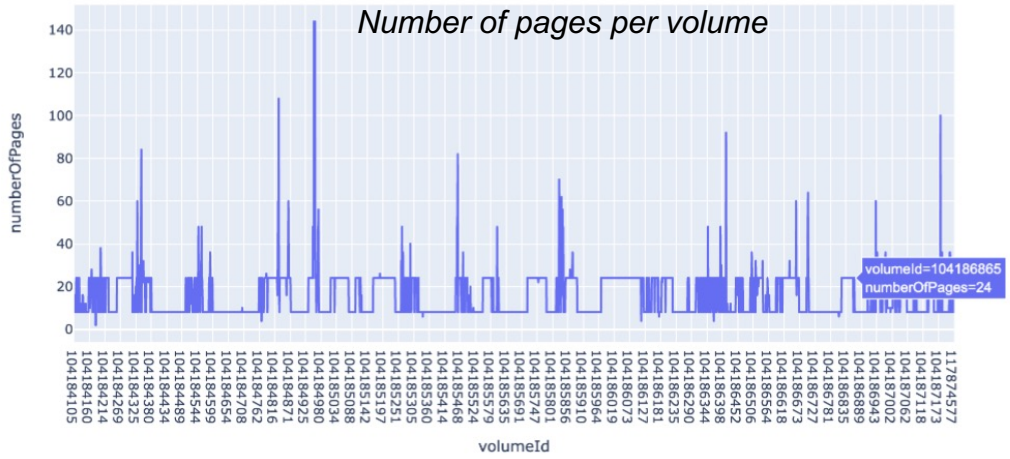
dataframe

MMSID	9937033633804341
edition	None
editor	Milne, John
editor_date	1792-1871
genre	Chapbooks-Scotland-Aberdeen-1801-1900
language	eng
metsXML	104184105-mets.xml
termsOfAddress	None
numberOfPages	8
numberOfWords	53
permanentURL	https://digital.nls.uk/104184105
physicalDescription	8 p. ; 18 cm.
place	Aberdeen
publisher	Printed by A. Imlay, 22, Long Acre
referencedBy	None
shelfLocator	L.C.2786.A(1)
altoXML	104184105/alto/107134030.34.xml
serieSubTitle	to the tune of Johnny Cop
text	A SONG JRAISB OP THE ^ HIGHLAND LADS. To the T...
pageNum	1
volumeTitle	song in praise of the highland lads
volumeId	104184105
year	1826
serieNum	0
part	0
collectionName	Chapbooks printed in Scotland
serieTitle	song in praise of the highland lads
publisherPersons	[]
numberOfVolumes	1
volumeNum	1

Name: 0, dtype: object

Chapbooks of Scotland





Phase 1.2: Information Extraction – Other NLS Collections

Frequency of genres per year. Each bubble indicates the frequency of a genre for a particular year

Applying advanced IE/AI/NLP to historical digital collections

St Andrews / CS

The Journey

Phase 1: Information Extraction

Phase 2: Ontologies and Knowledge Graphs:

- **2.1 - EB**
- **2.2 - NLS**

Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning

Phase 4: Defoe and Knowledge Graphs

Phase 5: React-Flask Web Platform

Phase 6: Case Studies

EB Ontology

St Andrews / CS

Aim: to capture a shareable and reusable knowledge representation of the EB.

EB Ontology → A formal description of knowledge as a set of concepts

In order to create and publish the EB-Ontology we 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

EB-Ontology available at:

- https://github.com/francesNLP/EB-ontology
- https://w3id.org/eb/

Phase 2.1: EB Ontology and Knowledge Graph

St Andrews / CS

EB Ontology

This version:
https://w3id.org/eb/0.0.2
Latest version:
https://w3id.org/eb
Revision:
0.0.2
Authors:
Rosa Filgueira
Contributors:
Daniel Garijo
Download serialization:
Format JSON LD | Format RDF/XML | Format N Triples | Format TTL
License:
License license name goes here
Visualization:
Visualize with WebVowl
Cite as:
Rosa Filgueira. EB Ontology. Revision: 0.0.2. Retrieved from: https://w3id.org/eb/0.0.2

Abstract

An ontology designed to represent volumes, terms and editions of the Encyclopaedia Britannica. The ontology extends Schema.org.

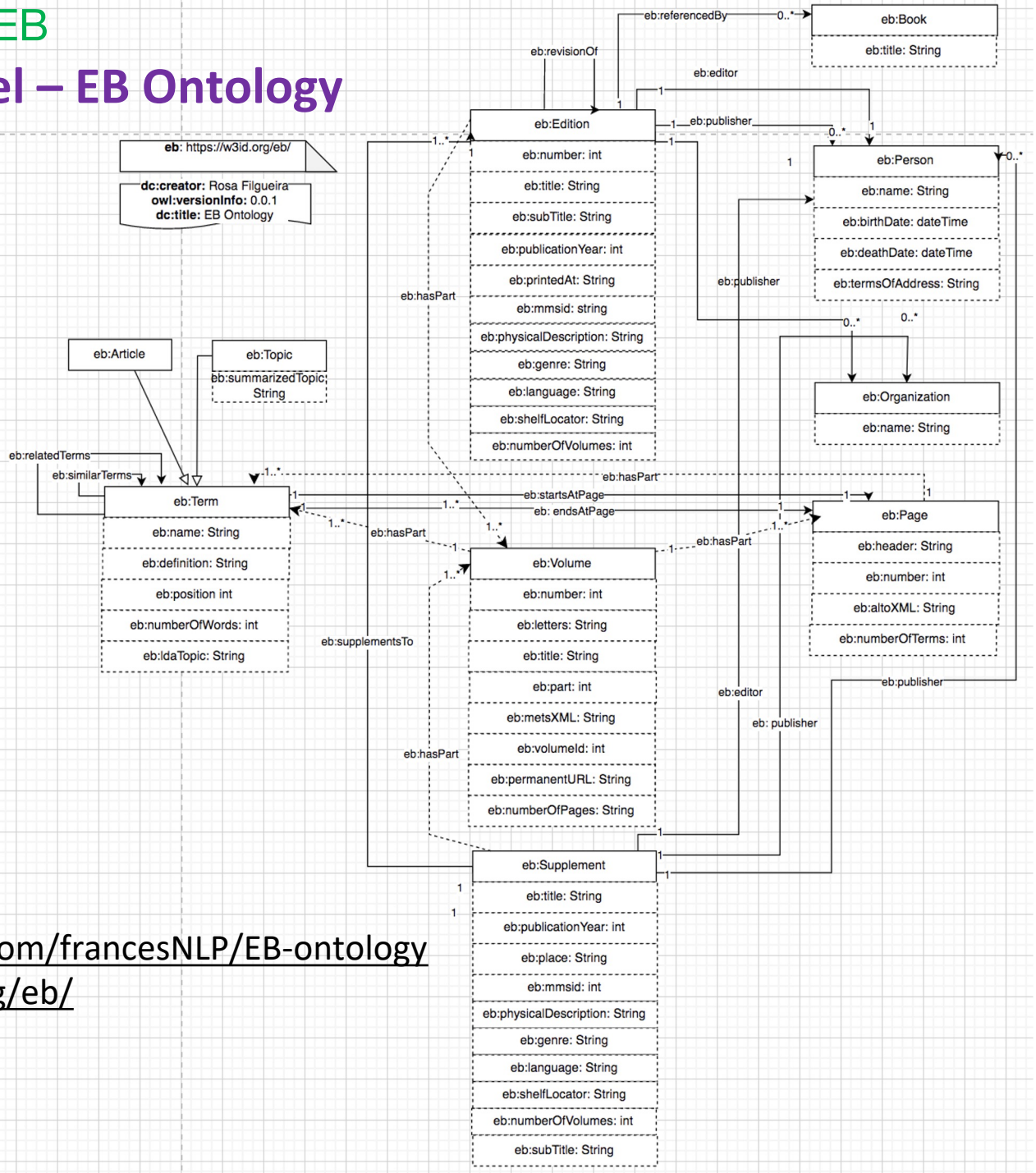
Table of contents

1. Introduction

This is a place holder text for the introduction. The introduction should briefly describe the ontology, its motivation, state of the art and goals.

Phase 2.1 - EB

Data Model – EB Ontology



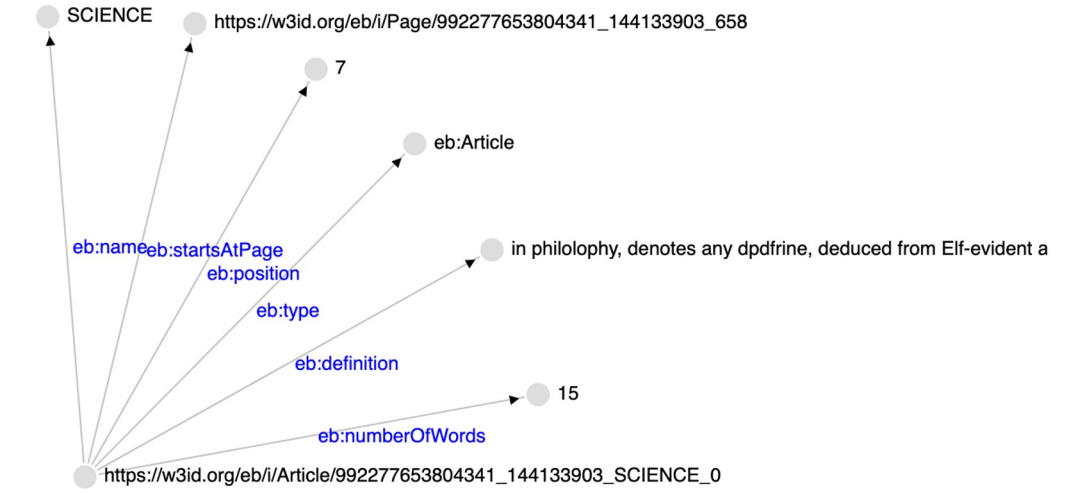
https://github.com/francesNLP/EB-ontology
https://w3id.org/eb/

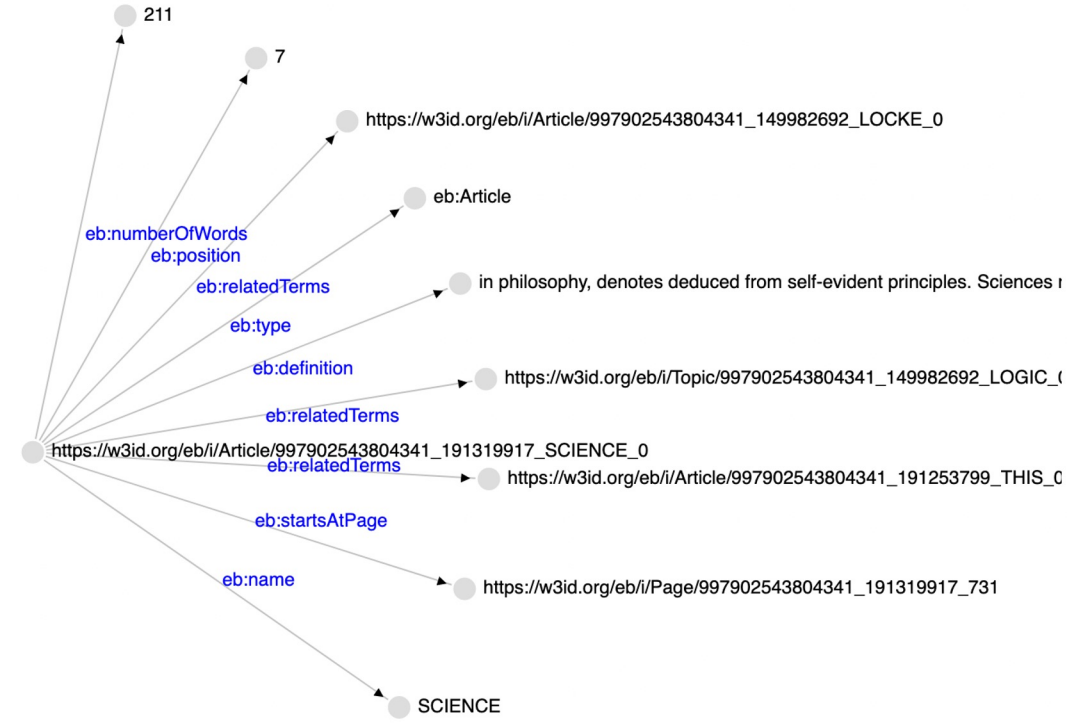
EB Knowledge Graph

St Andrews / CS

EB Knowledge Graph: Populating the post-processed information (extracted **Terms & Metadata**) into an RDF triple store + **EB Ontology**:
- Apache Jena FUSEKI SPARQL server

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





St Andrews / CS

- **Science** term in the First Edition issued in 1771

Edition 1 Year 1771
Edition 1 Year 1773
Edition 2 Year 1778
Edition 3 Year 1797
Edition 4 Year 1810
Edition 5 Year 1815
Edition 6 Year 1823
Edition 7 Year 1842
Edition 8 Year 1853

- **Science** term in the Third Edition issued in 1797

Phase 2.1: EB Ontology and Knowledge Graph

St Andrews / CS

zenodo Search Upload Communities

There is a **newer version** of this record available.

June 21, 2022 Dataset Open Access

EB-KG: Knowlege Graph of the first 8 eiditions Encyclopaedia Brittanica (1768-1860)

Rosa Filgueira

This Knowlege Graph represents the information of the first eight editions of Encyclopaedia Brittanica (years: 1768 to 1860) in RDF (ttl format).

The raw dataset is provided by the NLS in this link , and it comprises of eight editions and a total of 195 volumes with a total size of 44GB. It uses two XMLs schemas: METS for descriptive, structural, technical and administrative metadata (Title, Author, Publisher, etc); and ALTO for encoding the OCR text of a page.

In this work, we have extracted the information from METS and ALTO XMLS using defoe tool and developed novel information extraction heuristics. With the extracted information, we created the EB-KG Knowlege Graph, which uses the EB Ontolgy, to represent such information. Furthermore, during the information extraction phase, we have employed several techniques to mitigate two common OCR errors: long-S and the line-break hyphenation.

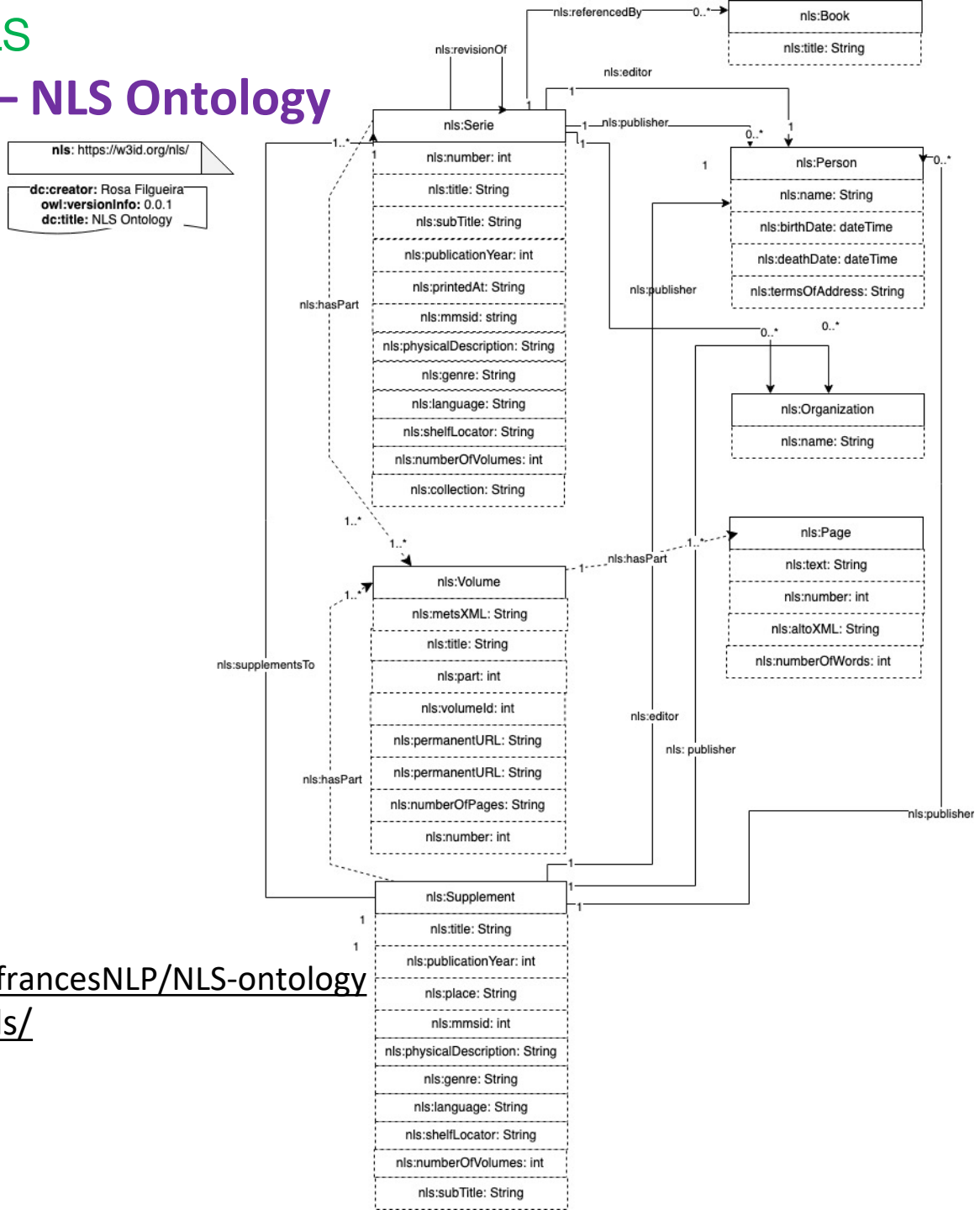
The EB-KG contains 1,638,239 RDF triples. It has information from 8 editions. Each edition can have several Volumes, references to Books, Supplements; it also has an Editor and a Publisher, which can be a Person or an Organization. A Volume has several Pages, which can contain several Terms. And a Term can be either a Topic (a term described across several pages, often combining text, pictures, and tables.) or an Article (a description of the term in one- or two-paragraph long text (similar to an entry in a dictionary)). The data model of the EB-KG can be found here.

The original ALTO files do not indicate the start and end of each EB term, the first part of our work involved the automated extraction of all terms (along with their metadata) across editions, so they can be analysed independently without the surrounding text.

This work was performed during my 2021-2022 National Library of Scotland Digital Scholarship Fellowship.

Phase 2.2 - NLS

Data Model – NLS Ontology



https://github.com/francesNLP/NLS-ontology

https://w3id.org/nls/

EB Knowledge Graph

St Andrews / CS

3 NLS Knowledge Graph: Populating the post-processed information into RDF triple store + **NLS Ontology**: Apache Jena FUSEKI SPARQL server

gazetters_scotland.ttl
chapbooks_scotland.ttl
ladies_debating.ttl

Phase 2.2: NLS Ontology and Knowledge Graphs

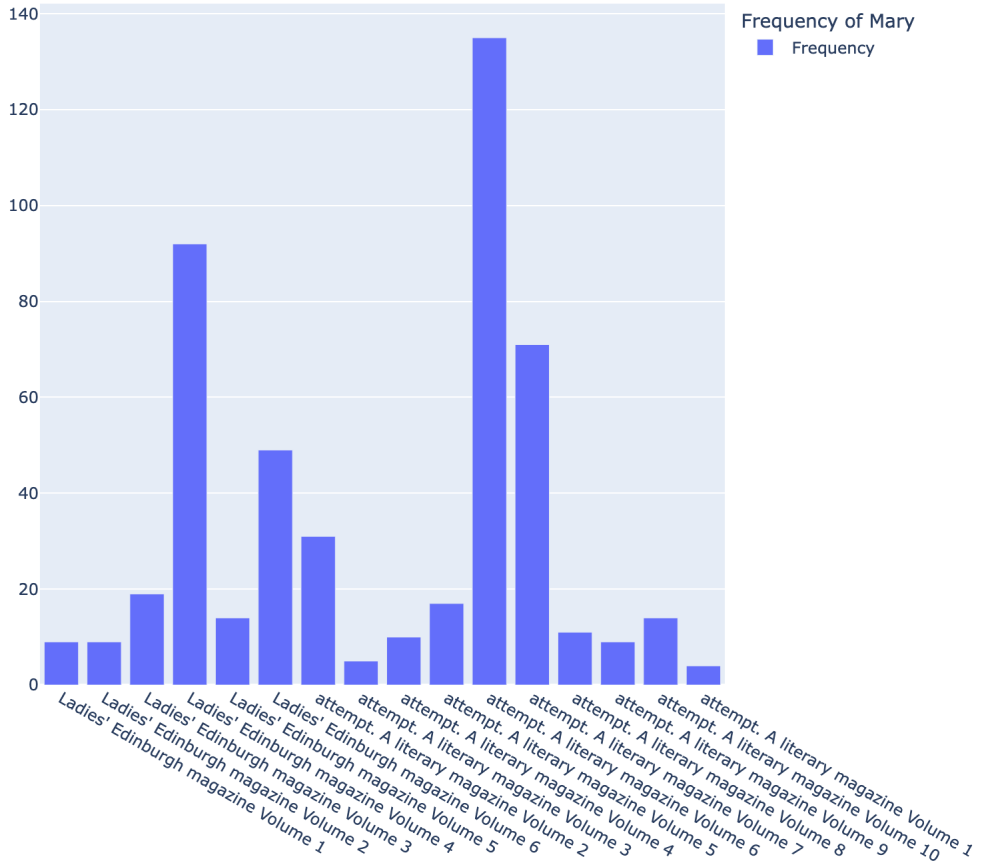
Calculating the frequency of the term ‘Mary’ across the publications of the *Ladies’ Edinburgh Debating Society* using ladies_debating.ttl

```
sparql = SPARQLWrapper("http://35.228.63.82:3030/ladies_debating/sparql")
sparql.setQuery("""
SELECT ?text ?title
        WHERE {
        ?page a nls:Page .
        ?page nls:text ?text .
        ?v nls:hasPart ?page .
        ?v nls:title ?title
        FILTER regex(?text, "Mary")
        }

""")
sparql.setReturnFormat(JSON)
results = sparql.query().convert()
```

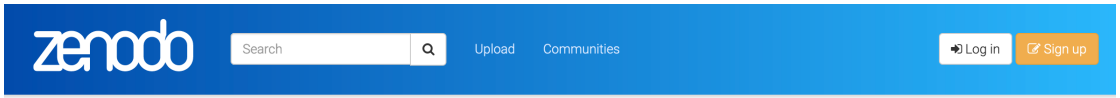
SPARQL Query to obtain the frequency of ‘Mary’ across volumes.

Visualization of the frequency results calculated above.



Phase 2.2: NLS Ontology and Knowledge Graphs

St Andrews / CS



June 22, 2022

Dataset Open Access

GazetteersScotland-KG: A Knowlege Graph for representing the Gazetteers of Scotland (1803-1901)

Rosa Filgueira

This Knowlege Graph represents the information of the "Gazeteers of Scotland" (years: 1803 - 1901) collection in RDF (ttl format). This collection comprises twenty volumes of the most popular descriptive gazetteers of Scotland in the 19th century. Principal places in Scotland, including towns, counties, castles, glens, antiquities and parishes, are listed alphabetically. Each entry includes detailed historical and geographical information about each place. The raw dataset is provided by the NLS in this link. As other NLS data collections, they are originally provided using two XMLs schemas: METS for descriptive, structural, technical and administrative metadata (Title, Author, Publisher, etc); and ALTO for encoding the OCR text of a page.

In this work, we have extracted the information from METS and ALTO XMLS using defoe tool and developed a new information extraction defoe query , and created a new Knowlege Graph called GazetteersScotland-KG. The GazetteersScotland-KG uses the NLS Ontology to represent the information extracted. Furthermore, during the information extraction phase, we have employed several techniques to mitigate two common OCR errors: long-S and the line-break hyphenation.

The GazetteersScotland-KG contains 354,998 RDF triples. It has information from 12 series and 20 volumes: Each serie can have several Volumes. Each serie has an Editorm Publisher, mmsid, Shelf-Locator, publication year, etc. A Volume has several Pages, with text in them. The data model of the GazetteersScotland-KG can be found here.

30 views

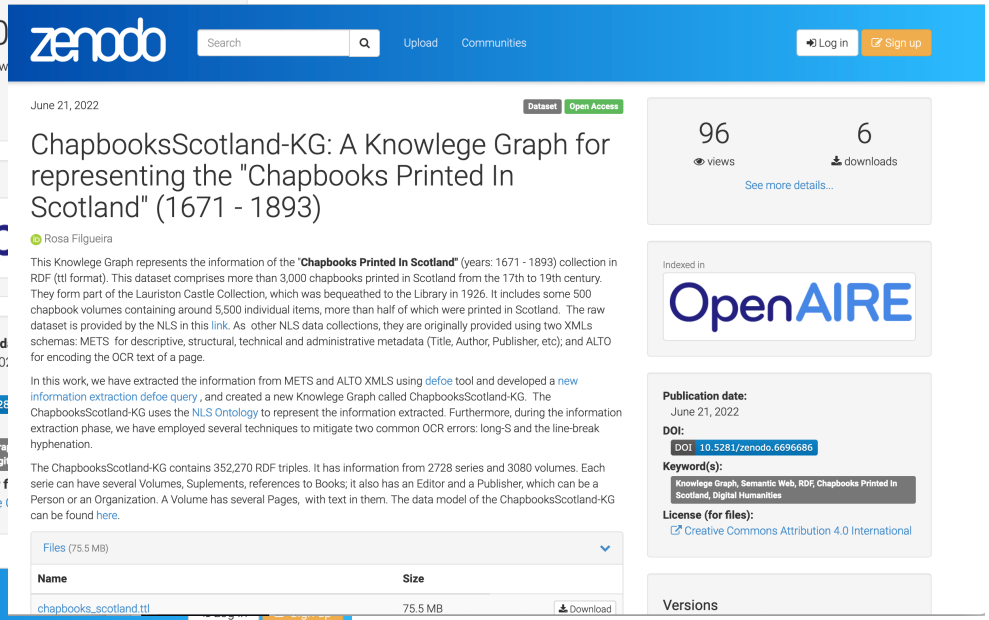
Indexed in Op

Publication d
June 22, 20

DOI:
DOI 10.525

Keyword(s):
Knowlege Gra
Scotland, Digi

License (for f
Creative (



June 21, 2022

Dataset Open Access

ChapbooksScotland-KG: A Knowlege Graph for representing the "Chapbooks Printed In Scotland" (1671 - 1893)

Rosa Filgueira

This Knowlege Graph represents the information of the "**Chapbooks Printed In Scotland**" (years: 1671 - 1893) collection in RDF (ttl format). This dataset comprises more than 3,000 chapbooks printed in Scotland from the 17th to 19th century. They form part of the Lauriston Castle Collection, which was bequeathed to the Library in 1926. It includes some 500 chapbook volumes containing around 5,500 individual items, more than half of which were printed in Scotland. The raw dataset is provided by the NLS in this link. As other NLS data collections, they are originally provided using two XMLs schemas: METS for descriptive, structural, technical and administrative metadata (Title, Author, Publisher, etc); and ALTO for encoding the OCR text of a page.

In this work, we have extracted the information from METS and ALTO XMLS using defoe tool and developed a new information extraction defoe query , and created a new Knowlege Graph called ChapbooksScotland-KG. The ChapbooksScotland-KG uses the NLS Ontology to represent the information extracted. Furthermore, during the information extraction phase, we have employed several techniques to mitigate two common OCR errors: long-S and the line-break hyphenation.

The ChapbooksScotland-KG contains 352,270 RDF triples. It has information from 2728 series and 3080 volumes. Each serie can have several Volumes, Suplements, references to Books; it also has an Editor and a Publisher, which can be a Person or an Organization. A Volume has several Pages, with text in them. The data model of the ChapbooksScotland-KG can be found here.

Files (75.5 MB)

Name	Size	
chapbooks_scotland.ttl	75.5 MB	Download

96 views 6 downloads

See more details...

Indexed in OpenAIRE

Publication date:
June 21, 2022

DOI:
DOI 10.5281/zenodo.6696686

Keyword(s):
Knowlege Graph, Semantic Web, RDF, Chapbooks Printed In Scotland, Digital Humanities

License (for files):
Creative Commons Attribution 4.0 International

Versions

June 22, 2022

Dataset Open Access

LadiesDebating-KG: A Knowlege Graph for representing the "Edinburgh Ladies' Debating Society Digital Collection" (1865 - 1880)

Rosa Filgueira

This Knowlege Graph represents the information of the "Edinburgh Ladies' Debating Society" (years: 1865 - 1880) collection in RDF (ttl format). This collection consists of the complete runs of two Edinburgh journals, **'The Attempt' (10 volumes, 1865-74)** and its successor **'The Ladies' Edinburgh Magazine' (6 volumes, 1875-80)**. These publications were produced by a leading Edinburgh women's club, known during the period as the Edinburgh Essay Society or the Ladies' Edinburgh Essay Society, but subsequently as the Ladies' Edinburgh Debating Society. The Society existed from 1865 to 1935. The raw dataset is provided by the NLS in this link. As other NLS data collections, they are originally provided using two XMLs schemas: METS for descriptive, structural, technical and administrative metadata (Title, Author, Publisher, etc); and ALTO for encoding the OCR text of a page.

In this work, we have extracted the information from METS and ALTO XMLS using defoe tool and developed a new information extraction defoe query , and created a new Knowlege Graph called LadiesDebating-KG. The LadiesDebating-KG uses the NLS Ontology to represent the information extracted. Furthermore, during the information extraction phase, we have employed several techniques to mitigate two common OCR errors: long-S and the line-break hyphenation.

The LadiesDebating-KG contains 38,279 RDF triples. It has information from 2 series and 16 volumes: **'The attempt'** serie has 10 volumes and **'The Ladies'** serie has 6 volumes . Each serie has an Editor, mmsid, Shelf-Locator, publication year, etc. A Volume has several Pages, with text in them. The data model of the LadiesDebating-KG can be found here.

43 views 2 downloads

See more details...

Indexed in OpenAIRE

Publication date:
June 22, 2022

DOI:
DOI 10.5281/zenodo.6686596

Keyword(s):
Knowlege Graph, Semantic Web, RDF, Chapbooks Printed In Scotland, Digital Humanities

License (for files):
Creative Commons Attribution 4.0 International

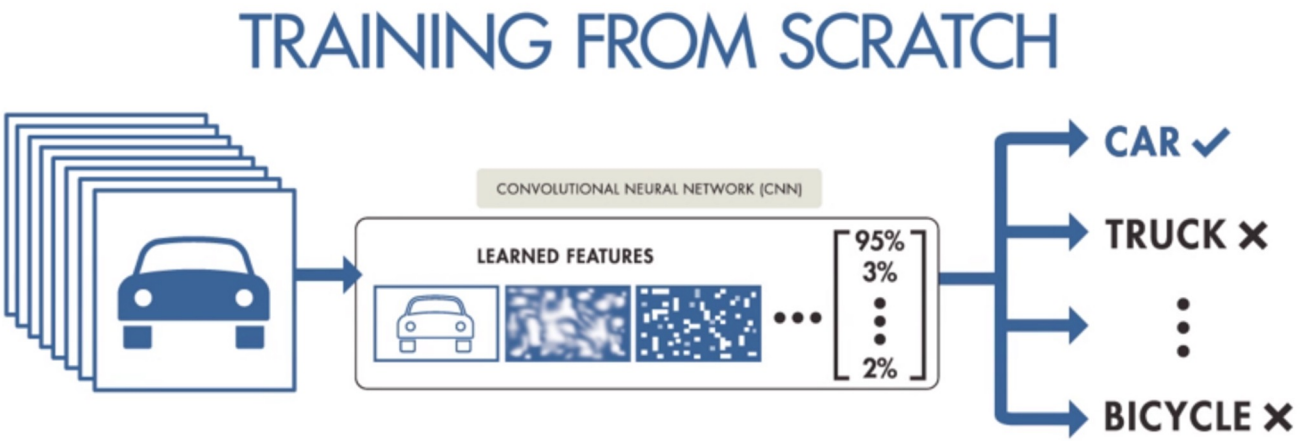
Applying advanced IE/AI/NLP to historical digital collections

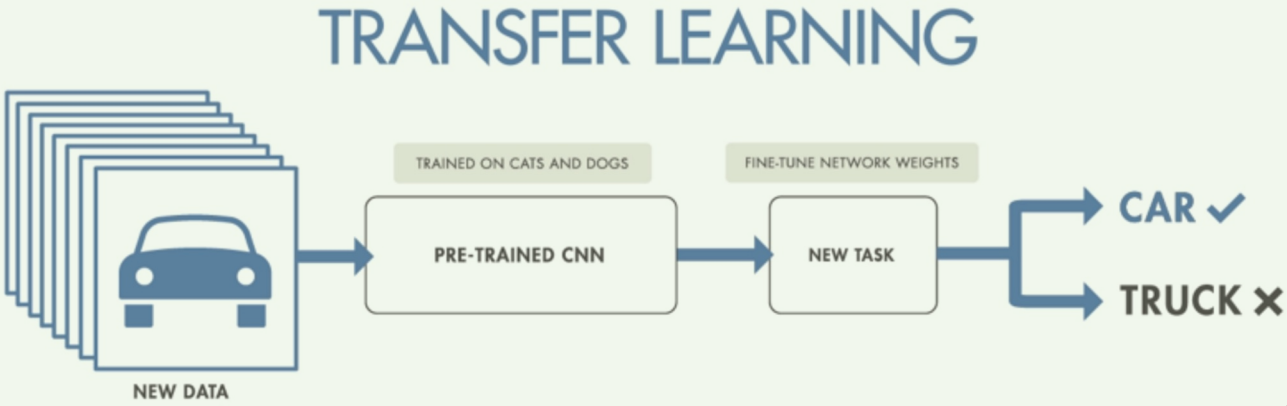
St Andrews / CS

The Journey

Phase 1: Information Extraction
Phase 2: Ontologies and Knowledge Graphs
Phase 3: Augmented Knowledge Graph with Deep Transfer Learning
Phase 4: Defoe and Knowledge Graphs
Phase 5: React-Flask Web Platform
Phase 6: Case Studies

Background





We are going to use this approach for applying NLP/ML analysis

Background

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.

Background

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.

Augmenting the EB-KG

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

- **sentiment analyses**: Classifying articles between positive and negative
  - transformer: *siebert/sentiment-roberta-large-english*
- **topic modelling**: Clustering terms into topics
  - transformer: *all-mpnet-base-v2*
- **term similarity**: Comparing terms & semantic similarity
  - transformer: *all-mpnet-base-v2*
- **spelling checking**: Finding misspelling/ocr errors and fixing them
  - transformer: *neuspell + ElmosclstmChecker*
- **summarization:** Summarizing the text of a topic term (XLNET)
  - transformer: *XLNeT*

Augmenting the EB-KG

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.

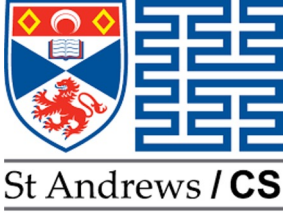
Cleaned Definition

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 efastern islands of Scotland , lying in 8\u00b0 and odd minutes W. long , and between 58\u00b0 and and 59 0 odd minutes N. latnd .

Applying advanced IE/AI/NLP to historical digital collections



The Journey

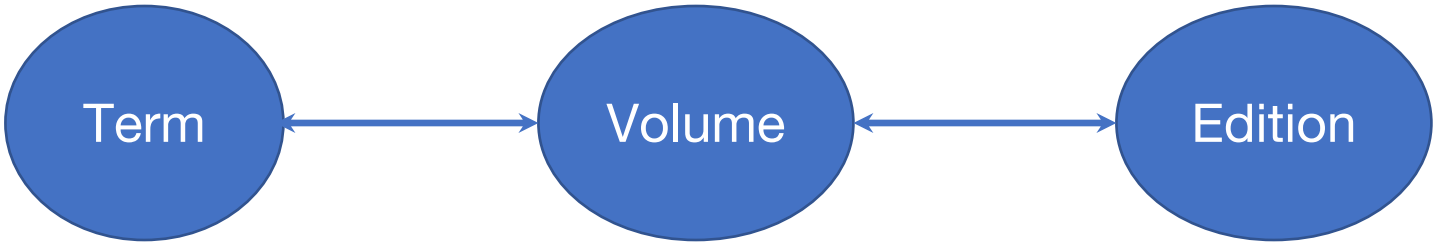
Phase 1**:** Information Extraction
Phase 2: Ontologies and Knowledge Graphs
Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning
Phase 4: Defoe and Knowledge Graphs
Phase 5: React-Flask Web Platform
Phase 6: Case Studies

St Andrews / CS

Querying the KGs

Two types of "queries":

- Type 1: **Extracting information from the EB Knowledge → SPARQL:** Navigating through the KG to get the desired information.



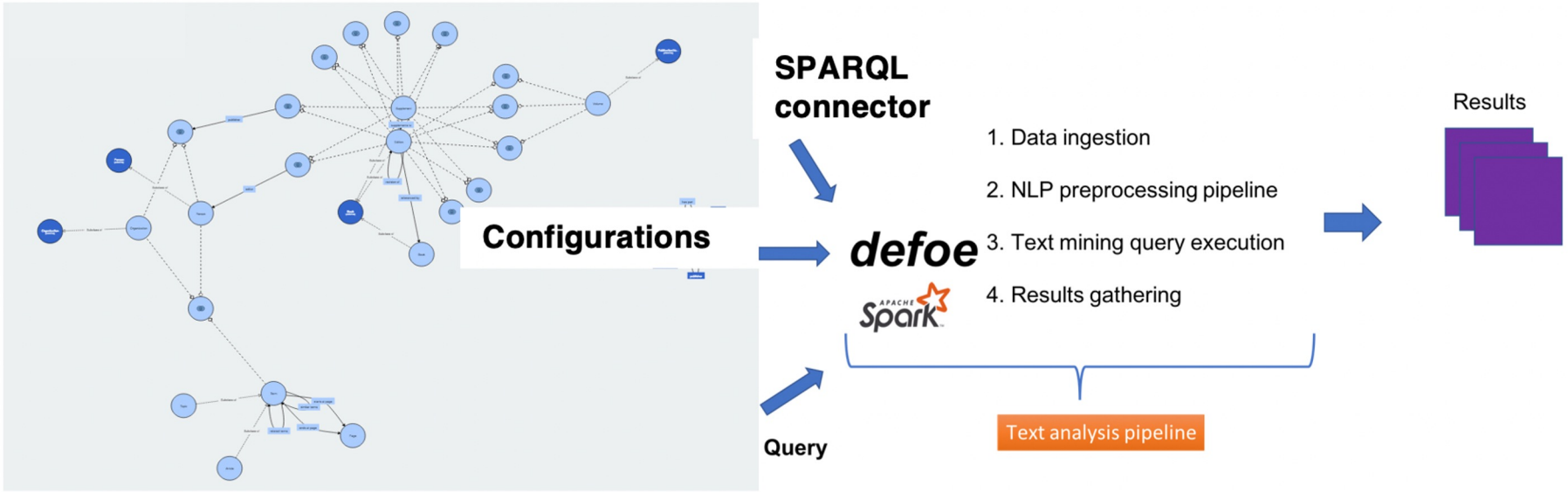
Example: Given a **Term** (e.g., *Edinburgh*), I can get the **Edition** Information (e.g., *Edition Title*)

- Type 2: **Processing information from the Knowledge Graphs →*defoe*:** Processing further the definitions from the selected terms/pages.
  - We needed to do some work on ***defoe*** first --> **Phase 4**

Phase 4: Defoe and Knowledge Graphs

St Andrews / CS

New **KG defoe connector (**based in SPARQL**)** to run ***defoe*** queries using the EB Knowledge Graph as a source of data → **defoe_KG**



Phase 4: Defoe and Knowledge Graphs

St Andrews / CS

New *defoe* queries fully configurable: different filtering options, target, lexicon, etc.

- *frequency_keysearch*: **Counts number of terms/pages** or **times** in which appear keywords or keysentences and group by years. Several filtering options.
- *fulltext_keysearch*: **Extracts terms definitions/pages** in which appear keywords or keysentences and group by years. Several filtering options.
- *snippet_keysearch*: **Extracts snippets of definitions/pages** in which appear keywords or keysentences and group by years. Filtering options, including the snippet size.
- *publication_normalization:* **Extracts the number of documents, pages, words** per year.
- *uris_keysearch:* **Extract uris of terms/pages** in which appear keywords or keysentences and group by years. Several filtering options.
- *geoparser_terms:* **Geoparsing the term definition/pages** in which appear keywords or keysentences and group by years. Several filtering options.
- *frequency-distribution:* Calculates the **frequency of the most 'N' common tokens** in terms definitions/pages.
- *lexicon-diversity*: Computes the **lexical diversity** metric for a given collection: ratio of the vocabulary size to the total number of words in the text.
- *person_entity_recognition*: **Identifies people mentioned** in text and estimates the gender distribution.

Applying advanced IE/AI/NLP to historical digital collections

St Andrews / CS

The Journey

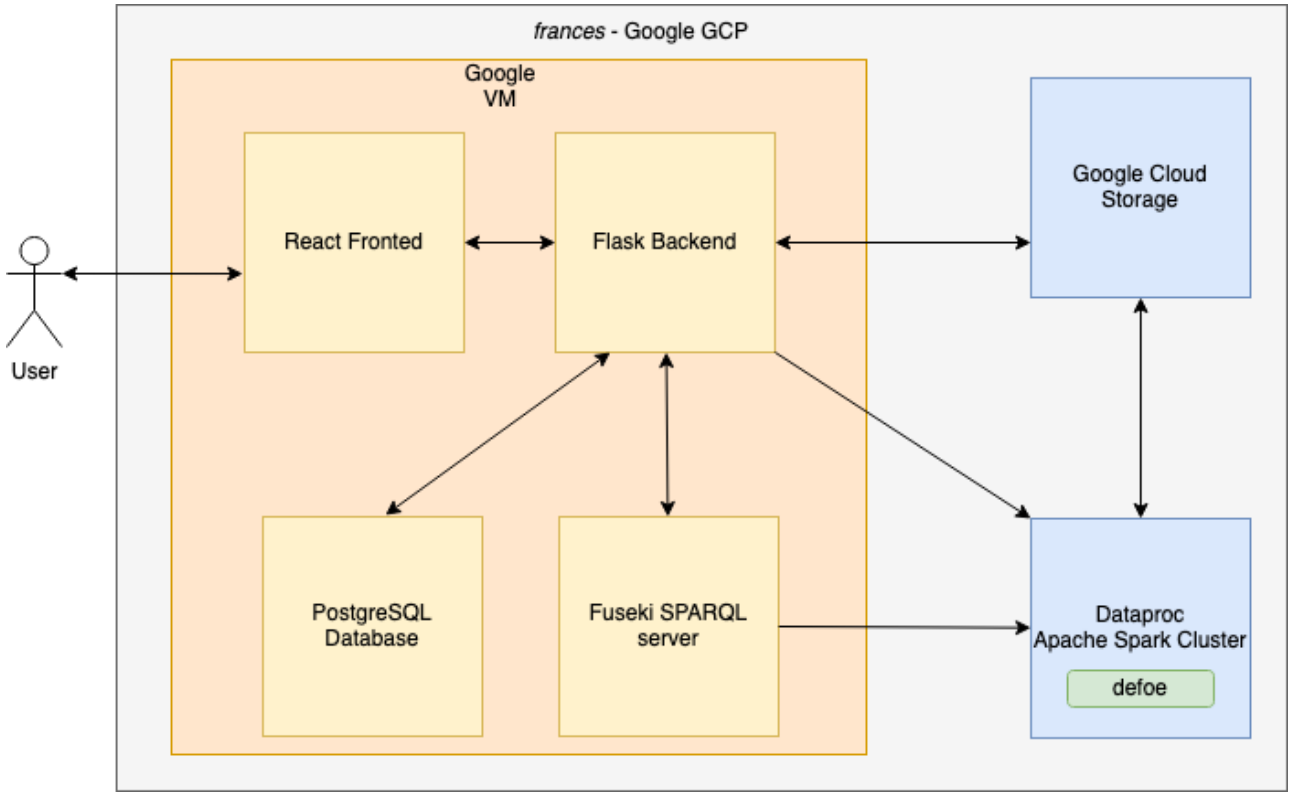
Phase 1: Information Extraction
Phase 2: Ontologies and Knowledge Graphs
Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning
Phase 4: Defoe and Knowledge Graphs
Phase 5: React-Flask Web Platform
Phase 6: Case Studies

Phase 5: React-Flask Web Platform

St Andrews / CS

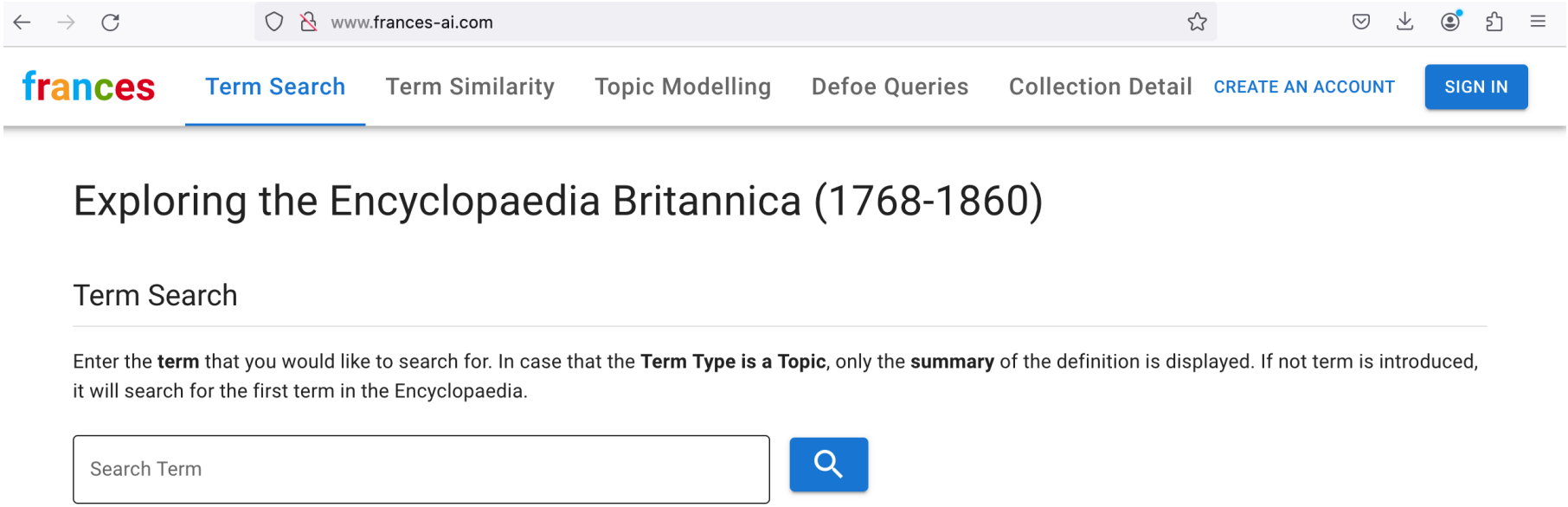
- Users DO NOT have to create SPARQL or ***defoe*** queries –> the web-application does it for them → Abstractions to SPARQL & DEFOE

- The web-app runs both types of queries and visualizes the results



Phase 5: React-Flask Web Platform

St Andrews / CS



Phase 5: React-Flask Web Platform

St Andrews / CS

frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | CREATE AN ACCOUNT | SIGN IN

Search Term: Flower

Results for **FLOWER**

Note that if you 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 id** if will take you to the **Topic Modelling page**, listing all the terms belonging to that particular topic model.

Year	Edition	Volume	Start Page	End Page	Term Type	Definition / Summary	Related Terms	Topic Modelling ID	Sentiment Score	Advanced Options
1771	1	2	566	566	Article	among botanifts and gardeners, the most beautiful part of trees and plants, containing the organs or parts of frutfification. See Botany,. Fr/cr W Flower. See Xeranthemum. Everlajiinx Flower.. See Gna...		3190	POSITIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS
1773	1	2	574	574	Article	among botaniffs and gardeners, the most beautiful part of trees and plants, containing the organs or parts of frutfification. See Botany., External Flower. See Xeranthemum, Everlafling Flower. See Gna...	BOTANY	3190	POSITIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS

1–2 of 9

Phase 5: React-Flask Web Platform

St Andrews / CS

frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | CREATE AN ACCOUNT | SIGN IN

Term Similarity

Enter **some text** that you would like to search similar terms for. If not term is introduced, it will search for the first term in the Encyclopaedia.

Place some text: Flower

The first **20** most similar results. The results are sorted by **similitud rank**.

Note that if you 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 id** if will take you to the **Topic Modelling page**, listing all the terms belonging to that particular topic model.

Year	Edition	Volume	Term	Definition	Topic Modelling ID	Similitud rank	Sentiment Score	Advanced Options
1823	6	8	FLOWER	Flos, among botanists and gardeners, the most beautiful part of trees and plants, containing the organs or parts of fructification. Seedesigned for medical use, should be jilnck— ■ / gj when they are ...	-1	0.5790331363677979	LABEL_0_0.87	VISUALISE CHECK SPELL SIMILAR TERMS
1823	6	16	PLENUS	FLOS, a full flower 5 a the highest degree of luxuriance in flowers. tany. Such flowers, although the the eye, are both vegetable monsters, to the sexualists, vegetable eunuchs; crease of the petals c...	1231	0.5691936016082764	LABEL_0_0.91	VISUALISE CHECK SPELL SIMILAR TERMS

Phase 5: React-Flask Web Platform

St Andrews / CS

frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | CREATE AN ACCOUNT | SIGN IN

Place some text: person who does scientific experiments

The first **20** most similar results. The results are sorted by **similitud rank**.

Note that if you 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 id** if will take you to the **Topic Modelling page**, listing all the terms belonging to that particular topic model.

Year	Edition	Volume	Term	Definition	Topic Modelling ID	Similitud rank	Sentiment Score	Advanced Options
1823	6	2	ARMORIST	a person skilled in the knowledge of	3715	0.5842932462692261	POSITIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS
1842	7	18	PROFESSOR	in the universities, a person who teaches or reads public lectures in some art or science. See Uni	441	0.5636317133903503	POSITIVE_0.99	VISUALISE CHECK SPELL SIMILAR TERMS
1815	5	16	PHYSICIAN	a person who profefies medicine, or the art of healing diseases. See Medicine. Pll YSI Cl AKS, College of, in London, Edinburgh, Dublin.	3251	0.5457887649536133	POSITIVE_1.00	VISUALISE CHECK SPELL

Phase 5: React-Flask Web Platform

St Andrews / CS

Topic Modelling Search

Enter a **topic name** or just the **ID** of a topic model to see all the terms within the same topic. All topics modelling names start with a number. If not topic is introduced. it will use the first topic modelling, which name starts with '0'.

Topic Name or ID: 3715

11 terms found for the topic**3715_knowledge_acquired_any_teacher_use_without_teacher_mr**

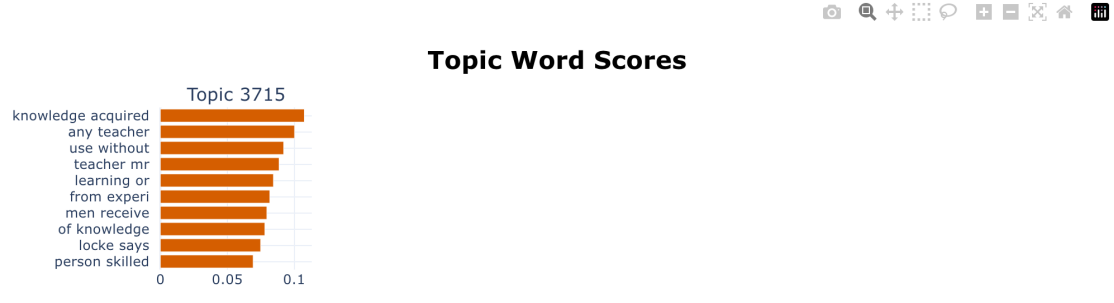
Note that if you instead click over a **term**, it will take you to the **Term Search page**, showing all the searching results for that term.

Year	Edition	Volume	Term	Definition	Sentiment Score	Advanced Options
1810	4	16	PHILOMATHES	a lover of learning or fcieace. thes,	POSITIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS
1823	6	2	ARMORIST	a person skilled in the knowledge of	POSITIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS
1815	5	2	ARMORIST	a person Ikilled in the knowledge of	NEGATIVE_1.00	VISUALISE CHECK SPELL SIMILAR TERMS

Topic Modelling Visualisations

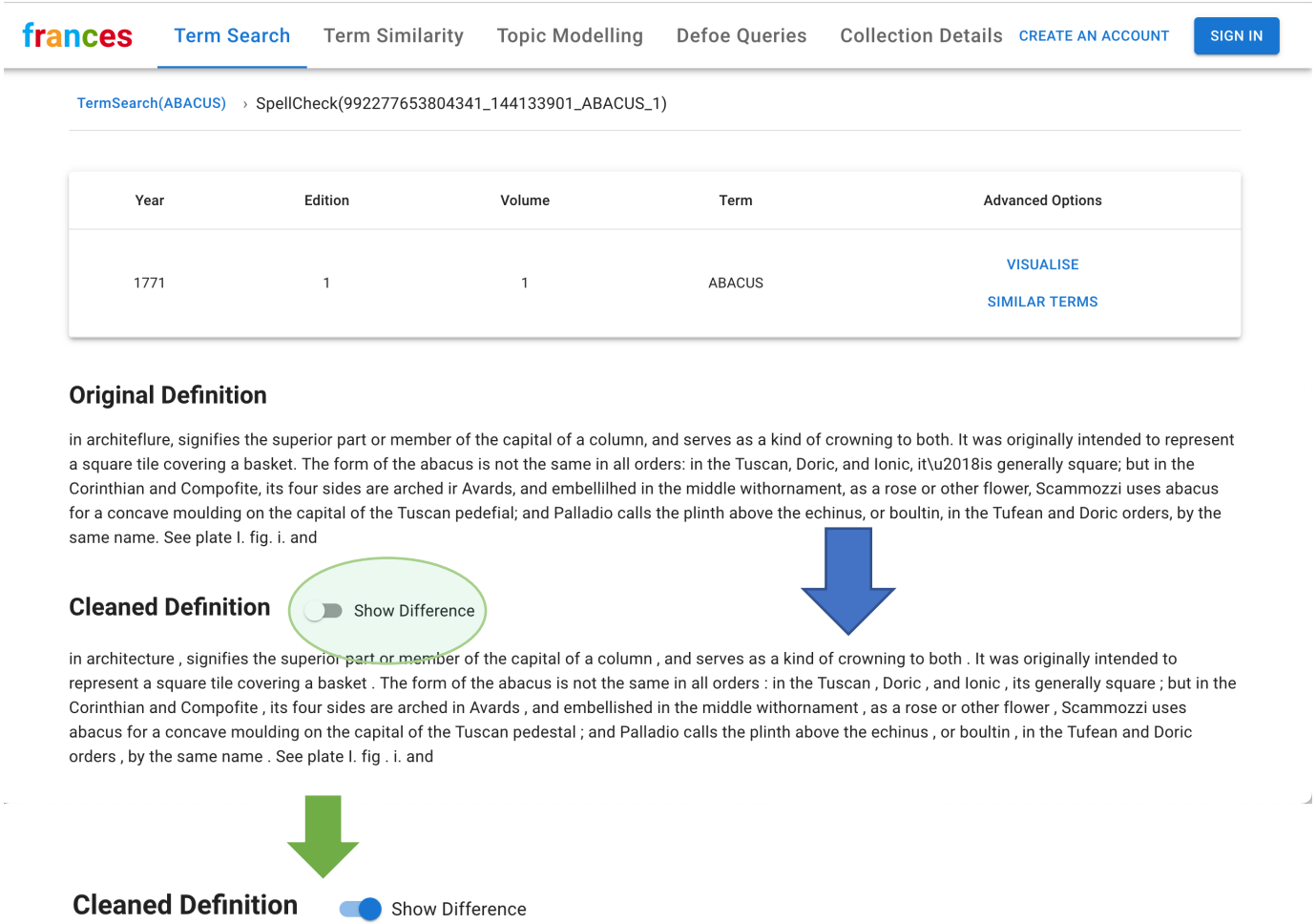
We are visualising each topic model found in our term-search. For each topic model, we plot their most common 10 words , along with their c-TF-IDF scores.

Note that the topic that starts with "-1" refers to all outliers and should typically be ignored



Phase 5: React-Flask Web Platform

St Andrews / CS

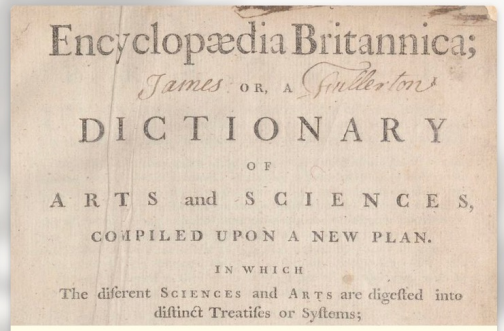


Phase 5: React-Flask Web Platform

St Andrews / CS

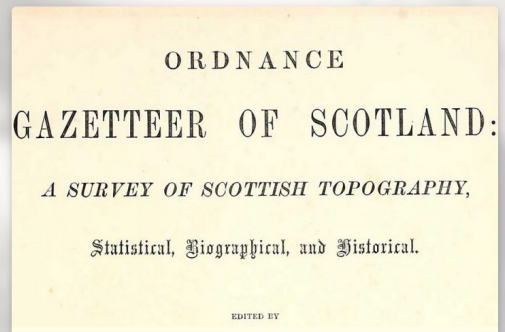
frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | CREATE AN ACCOUNT | SIGN IN

Exploring NLS Digital Collections



Encyclopaedia Britannica,1768-1860

Chapbooks printed in Scotland,1700-1899

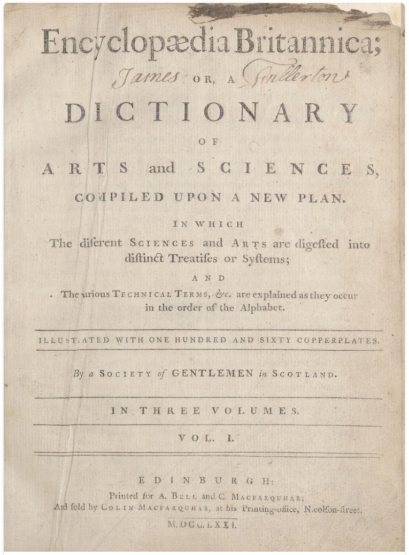


Gazetteers of Scotland,1803-1901

Ladies' Edinburgh Debating Society,1865-1880

Phase 5: React-Flask Web Platform

St Andrews / CS



Covers years	1768-1860
Number of image files	155388

SOURCE

Encyclopaedia Britannica

The first eight editions of Encyclopaedia Britannica, issued from 1768-1860, comprise a total of 143 volumes. The Britannica was first issued in Edinburgh in 100 weekly parts (forming 3 volumes) from 1768 to 1771 and illustrated with 160 copperplate engravings. The enterprise was undertaken by the partnership of printer Colin Macfarquhar (1744-1793) and engraver Andrew Bell (1726-1809) who paid William Smellie (1740-1795) to compile the Britannica's first edition for a fee of £200.

Subsequent editions of the Britannica expanded the content: the second edition was published in 10 volumes (1777-1784); the third in 18 volumes (1788-1797). As the Britannica expanded it sought contributions from leading experts in their fields who were either approached by the editors or drawn by the encyclopaedia's growing reputation. Macfarquhar and Bell retained the copyright for the first three editions before publication was taken over by Archibald Constable and then A & C Black. Managed and published in Edinburgh up to the 9th edition.

The Britannica set the standard for modern encyclopedias and is sometimes seen as an enduring product of the Scottish Enlightenment. These volumes were a compendium of current and practical knowledge made relatively affordable by the initial efforts of Macfarquhar and Bell and by Smellie's views on the democratisation of knowledge and the axiom, with which he opens the Preface to the first edition, that "utility ought to be the principal intention of every publication."

Editions for Encyclopaedia Britannica

Search Edition

- Edition 1,1771
- Edition 1,1773
- Edition 2,1778
- Edition 3,1797
- Sup. Edition 3, 1801
- Edition 4,1810
- Edition 5,1815

Title:	Edition 1,1771
SubTitle:	Illustrated with one hundred and sixty copperplates
Publication Year:	1771
Number:	1
Printed At:	Edinburgh
Shelf Locator:	EB.1
Physical Description:	3 v., 160 plates : ill. ; 26 cm. (4to)
MMSID	992277653804341
Genre:	encyclopedia
Language:	English
Number of Volumes:	3

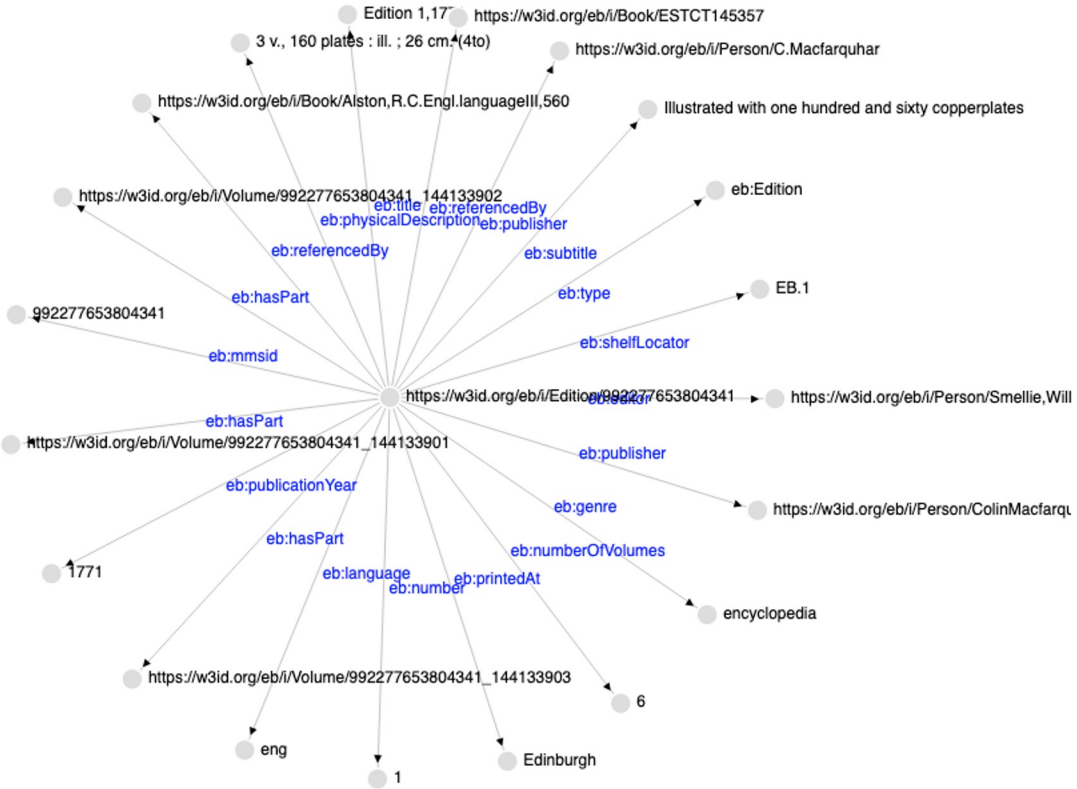
Volumes for Edition 1,1771

Search Volume

- 2 C-L
- 3 M-Z
- 1 A-B

Title:	Encyclopaedia Britannica; or, A dictionary of arts and sciences, compiled upon a new plan
Permanent URL:	https://digital.nls.uk/144133902
ID:	144133902
Number:	2
Number of Pages:	1018
Letters:	C-L

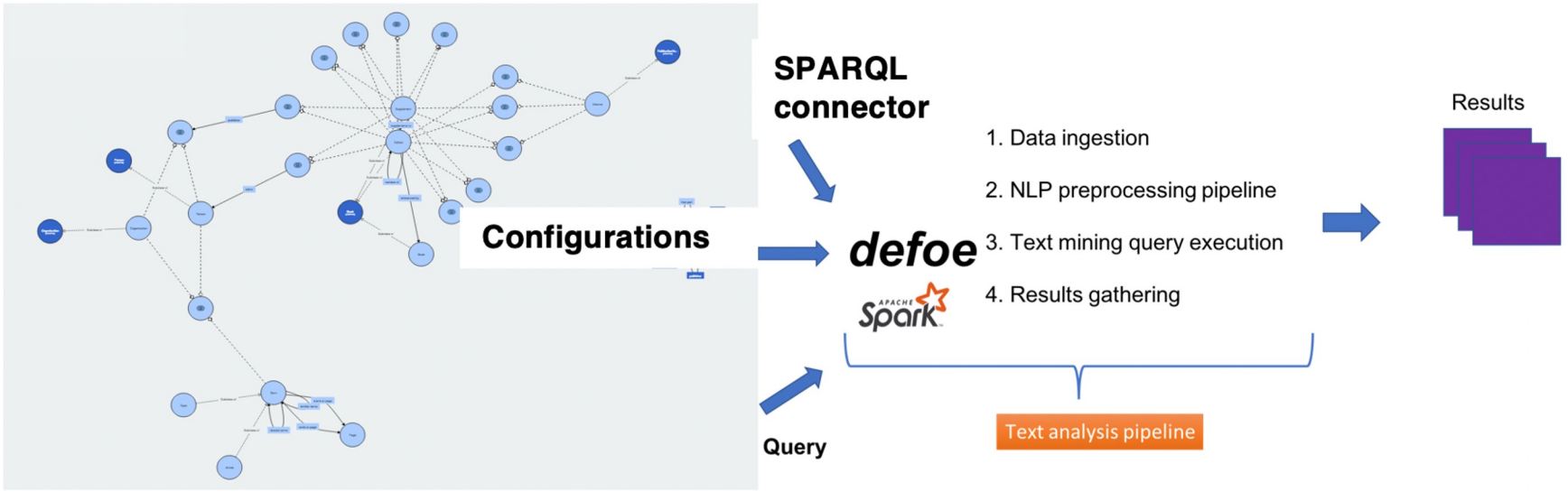
Visualization of Resources



Visualization of the First Edition (issued in 1771) resource stored in the EB-KG

St Andrews / CS

Defoe Queries – Text Mining Platform




```
def do_query(archives, config_file=None, logger=None, \
                                          context=None):
    # [archive, archive, ...]
    documents = archives.flatMap(lambda archive:\
                                 list(archive))
    # [num_pages, num_pages, ...]
    num_pages = documents.map(lambda document: \
                                document.num_pages)
    result = [documents.count(), num_pages.reduce(add)]
    return {"num_volumes": result[0],\
            "num_pages": result[1]}
```

```
Result:
num_volumes: 10
num_pages: 9448
```

Results using the 2nd Edition – EB

total_pages query: Iterates through archives and counts the total number of documents (e.g. volume, book etc) and total number of pages.

Phase 5: React-Flask Web Platform

St Andrews / CS

frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | DY

Defoe Queries

Collection: Encyclopaedia Britannica

Query type: frequency_keysearch_by_year

It counts the number of terms/words in which appear your selected kewyords/keysentences. It groups results by years.

- Encyclopaedia Britannica
- Chapbooks printed in Scotland
- Gazetteers of Scotland
- Ladies' Edinburgh Debating Society

t to the text.

None

Hit Count

It counts the number of times a keyword/keysentece appears in a term (article or topic).

Word Term

Upload Lexicon File

The file should contain a line per keyword and/or key sentence that you want to use in your query.

UPLOAD FILE

Filtering Options (Optional)

Target words or sentences

Word or Sentence ADD

Select term which contains **Any** the target words/sentences

Or contains **All** of the target words/sentences.

Year Range

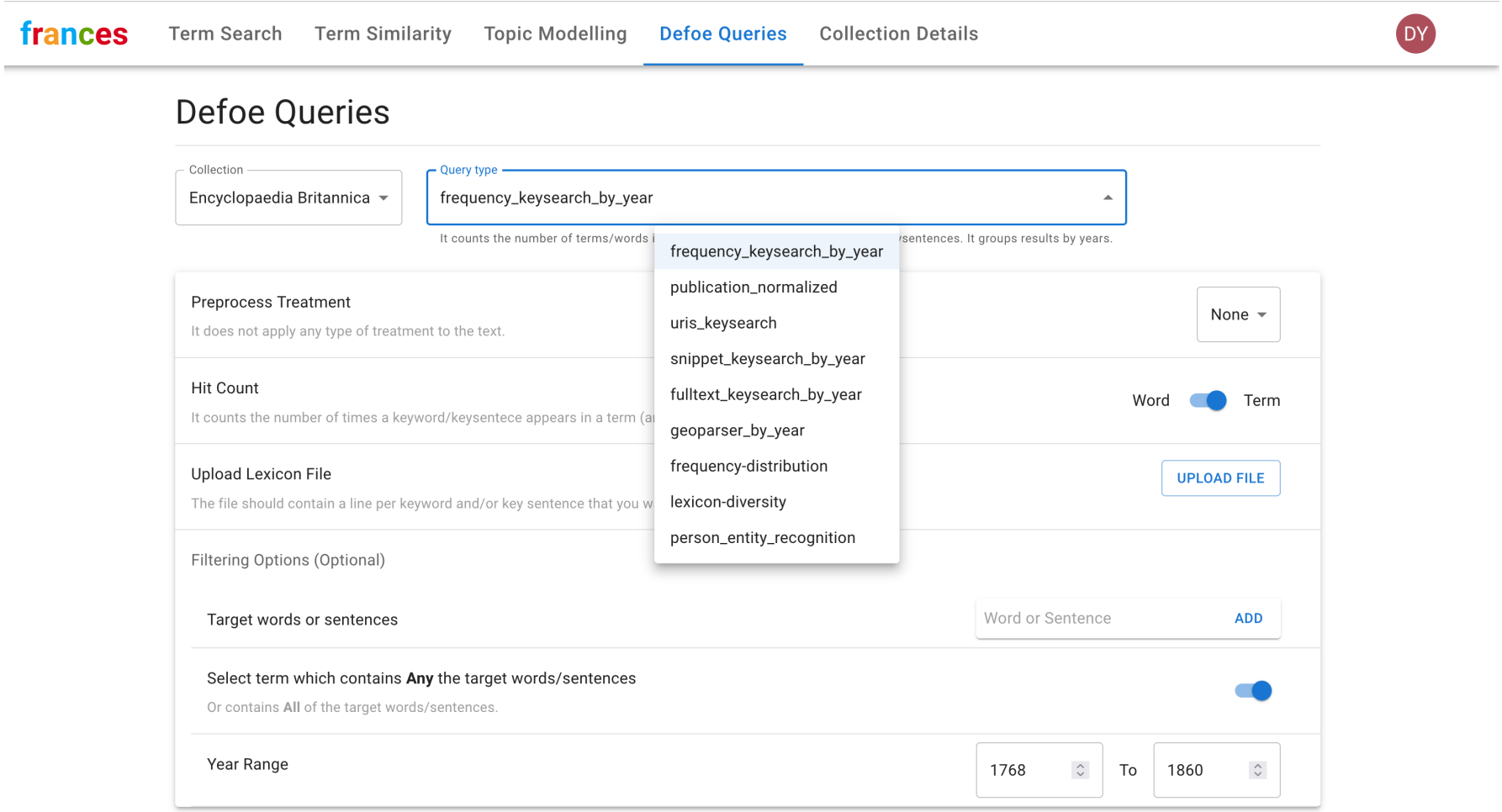
1768 To 1860

SUBMIT QUERY

Phase 5: React-Flask Web Platform

Defoe Queries – Text Mining Platform

St Andrews / CS



Phase 5: React-Flask Web Platform

Defoe Queries – Text Mining Platform

St Andrews / CS

frances | Term Search | Term Similarity | Topic Modelling | Defoe Queries | Collection Details | DY

- Profile
- Saved
- Tasks
- Sign out

Defoe Query Tasks

Collection	Query Type	Configuration	State	Submit Time	Actions
Chapbooks printed in Scotland	geoparser_by_year	animal.txt	DONE	2023-05-31 20:02:51.880409	VIEW
Encyclopaedia Britannica	frequency_keysearch_by_year	commodities.txt lemmatize term	DONE	2023-06-13 18:41:38.677508	VIEW
Ladies' Edinburgh Debating Society	publication_normalized		DONE	2023-06-13 18:42:10.114119	VIEW
Chapbooks printed in Scotland	geoparser_by_year	scots.txt lemmatize	DONE	2023-06-14 11:55:08.393946	VIEW
Chapbooks printed in Scotland	geoparser_by_year	scots.txt normalize	DONE	2023-06-14 13:10:14.124065	VIEW
Chapbooks printed in Scotland	geoparser_by_year	scots.txt	DONE	2023-06-20 08:35:23.817690	VIEW
Chapbooks printed in Scotland	geoparser_by_year	scots.txt lemmatize	DONE	2023-06-20 08:47:17.244466	VIEW
Encyclopaedia Britannica	publication_normalized		DONE	2023-06-21 20:42:21.304656	VIEW
Chapbooks printed in Scotland	publication_normalized		DONE	2023-06-21 20:42:36.779318	VIEW

Applying advanced IE/AI/NLP to historical digital collections

St Andrews / CS

The Journey

Phase 1**:** Information Extraction
Phase 2: Ontology and Knowledge Graphs
Phase 3: Augmented Knowledge Graphs with Deep Transfer Learning
Phase 4: Defoe and Knowledge Graphs
Phase 5: React-Flask Web Platform
Phase 6: Case Studies
- **6.1 Commodity and Enslaved Labor in the EB**
- **6.2 Geoparsing the Chapbooks printed in Scotland**

Phase 6.1: Commodity and Enslaved Labor in the EB

St Andrews / CS

Collaboration with: Ash Charlton & Melissa Terras

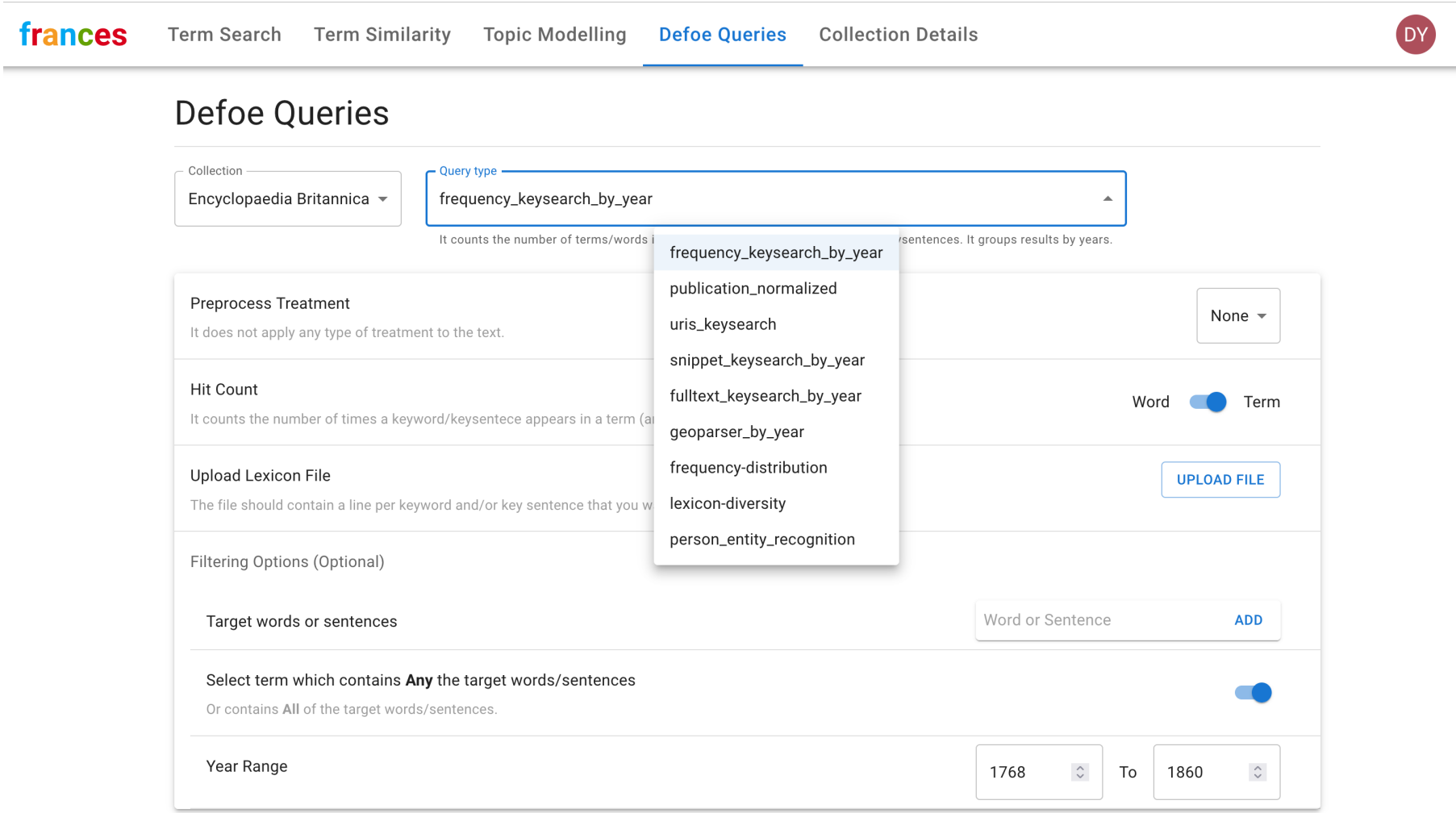
This case study examines how race, slavery, and products made by enslaved labor were portrayed in the Encyclopaedia Britannica (EB) during a time of intellectual and industrial growth.

This case study conducts an analysis of references to these topics, with a specific focus on **cotton**.

Phase 6.1: Commodity and Enslaved Labor in the EB

St Andrews / CS

Collaboration with: Ash Charlton & Melissa Terras



Phase 6.1: Commodity and Enslaved Labor in the EB

St Andrews / CS

Collaboration with: Ash Charlton & Melissa Terras

Collection: Encyclopaedia Britannica

Query type: frequency_keysearch_by_year

It counts the number of terms/words in which appear your selected kewyords/keysentences. It groups results by years.

Preprocess Treatment

It normalizes the text first and lemmatizes it later, returning the base (lemma) of each word.

Normalize & Lemmatize

Hit Count

It counts the number of times a keyword/keysentece appears in a term (article or topic).

Word — Term

Upload Lexicon File

The file should contain a line per keyword and/or key sentence that you want to use in your query.

UPLOAD FILE

20230723-142924_commodities.txt

Filtering Options (Optional)

Target words or sentences: Word or Sentence — ADD

Select term which contains **Any** the target words/sentences

Or contains **All** of the target words/sentences.

Year Range: 1768 To 1860

SUBMIT QUERY

Commodities associated with slavery: ‘tobacco’, ‘rum’, ‘indigo’, ‘sugar’, coffee’, ‘rice’, ‘cotton’ and ‘molasses’

Phase 6.1: Commodity and Enslaved Labor in the EB

Defoe Queries

Collaboration with: Ash Charlton & Melissa Terras

St Andrews / CS

Collection: Encyclopaedia Britannica

Query Type: frequency_keysearch_by_year

Lexicon Filename: commodities.txt

Preprocess: Normalize & Lemmatize

Hit Count: Term

Year range: 1768 - 1860

Submitted time: 2023-05-16 12:29:34.445029

CREATE ANOTHER QUERY | CHECK ALL QUERY TASKS

Result

DOWNLOAD

Year	Lexicon	Frequency
1771	sugar	35
	coffee	3
	indigo	13
	rice	12
	cotton	31
	tobacco	15
	rum	9
1773	rice	9
	cotton	33
	indigo	16

Rows per page: 10 1–10 of 76

Phase 6.1: Commodity and Enslaved Labor in the EB

Defoe Queries

Collaboration with: Ash Charlton & Melissa Terras

St Andrews / CS

Collection: Encyclopaedia Britannica

Query Type: frequency_keysearch_by_year

Lexicon Filename: commodities.txt

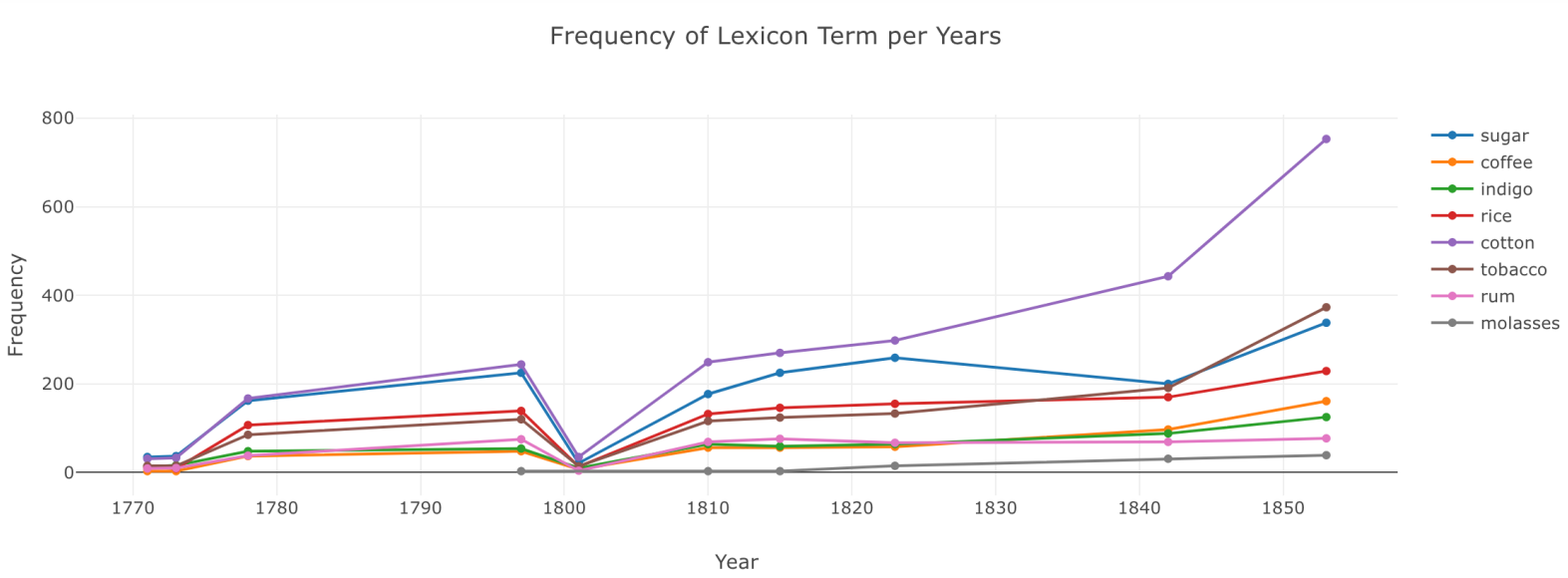
Preprocess: Normalize & Lemmatize

Hit Count: Term

Year range: 1768 - 1860

Submitted time: 2023-05-16 12:29:34.445029

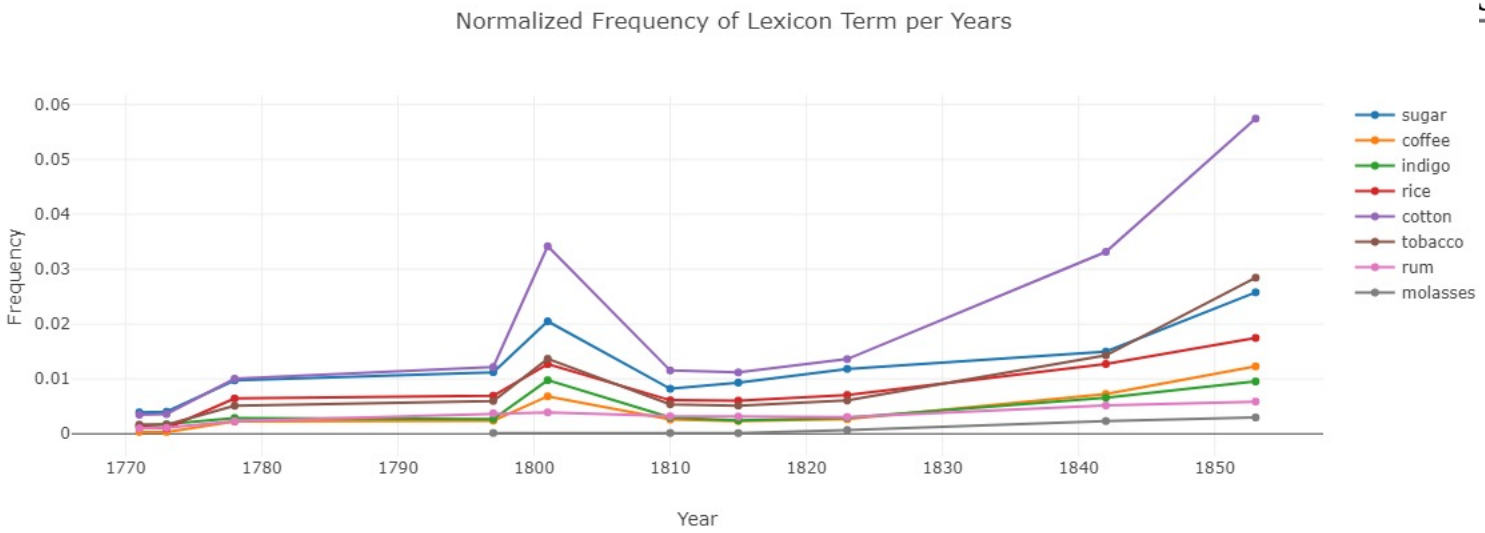
CREATE ANOTHER QUERY | CHECK ALL QUERY TASKS



N-gram to visualise the frequencies of eight commodities

Phase 6.1: Commodity and Enslaved Labor in the EB

Collaboration with: Ash Charlton & Melissa Terras



N-gram to visualise the normalized frequencies of eight commodities

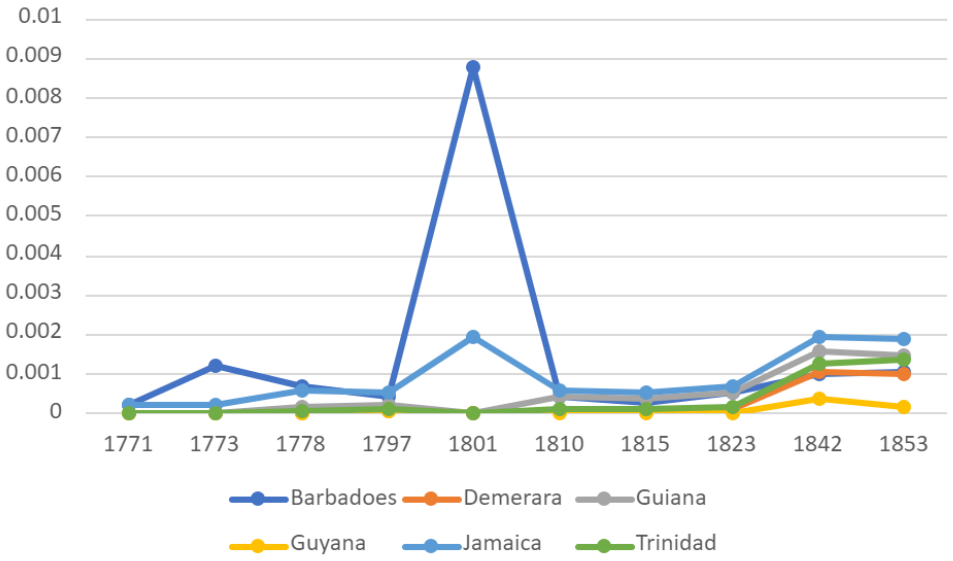
The analysis uncovered shifts in the frequency of commodity mentions over time, with '***cotton' surpassing 'sugar' as the most referenced commodity after 1801***.

The notable increase in terms mentioning cotton can be attributed to Britain's ***advancements in cotton technologies during the late eighteenth and nineteenth centuries***, facilitating mass production and manufacturing

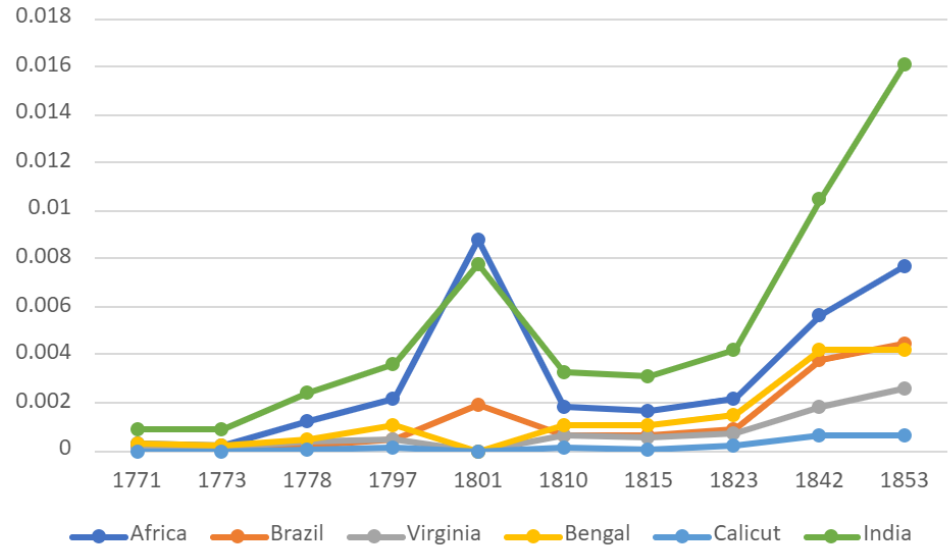
Phase 6.1: Commodity and Enslaved Labor in the EB

St Andrews / CS

Collaboration with: Ash Charlton & Melissa Terras



Occurrences of ‘cotton’ in EB terms mentioning **Caribbean locations**: ‘Barbadoes’, ‘Demerara’, ‘Guiana’, ‘Guyana’, ‘Jamaica’ and ‘Trinidad’

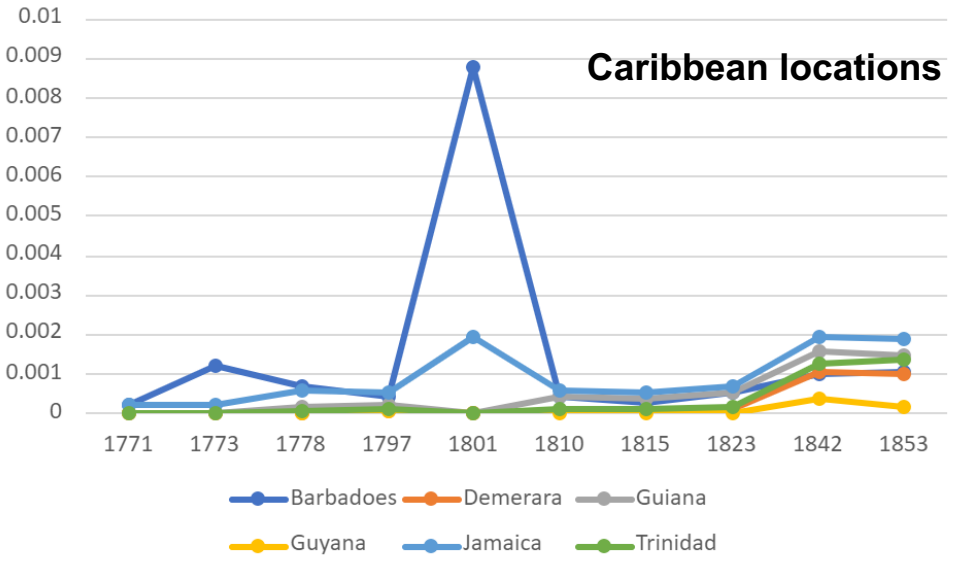


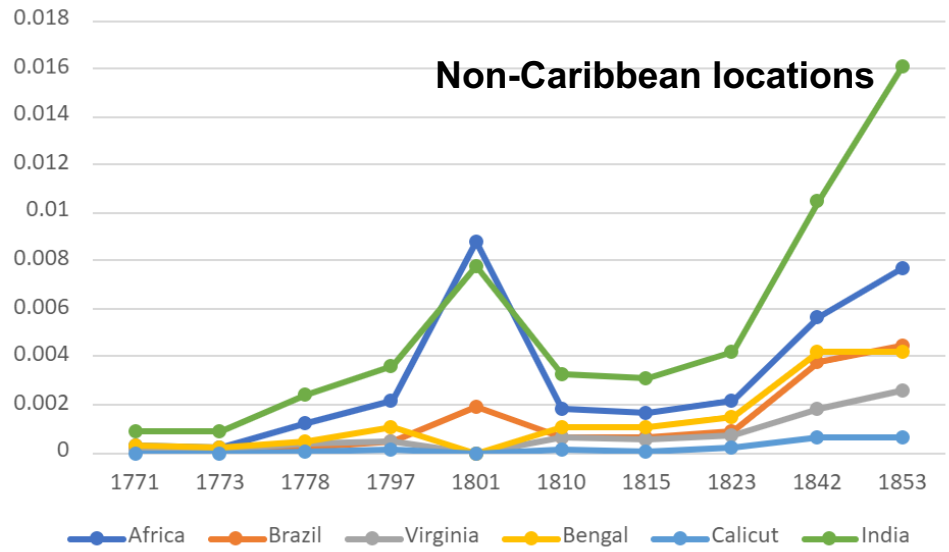
Occurrences of ‘cotton’ in EB terms mentioning **non-Caribbean** places: ‘Africa’, ‘Brazil’, ‘Virginia’, ‘Bengal’, ‘Calicut’, ‘India

Phase 6.1: Commodity and Enslaved Labor in the EB

St Andrews / CS

Collaboration with: Ash Charlton & Melissa Terras





The analysis shows a shift in cotton references from the Caribbean to India after 1823, likely due to Britain's East India Company's trading activities.

Phase 6.2: Geoparsing the Chapbooks printed in Scotland

St Andrews / CS

Defoe Queries

Collection: Chapbooks printed in Scotland

Query type: geoparser_by_year

It geo-locates locations in pages and geo-resolves them using the Edinburgh Geoparser. It groups results by years.

Preprocess Treatment

It does not apply any type of treatment to the text.

None

None
Normalize
Normalize & Numbers
Normalize & Lemmatize
Normalize & Stemming

Upload Lexicon File

The file should contain a line per keyword and/or key sentence that you want to use in your query.

Gazetteer

A world-wide gazetteer of over eight million placenames, made available free of charge.

Filtering Options (Optional)

Target words or sentences

Word or Sentence ADD

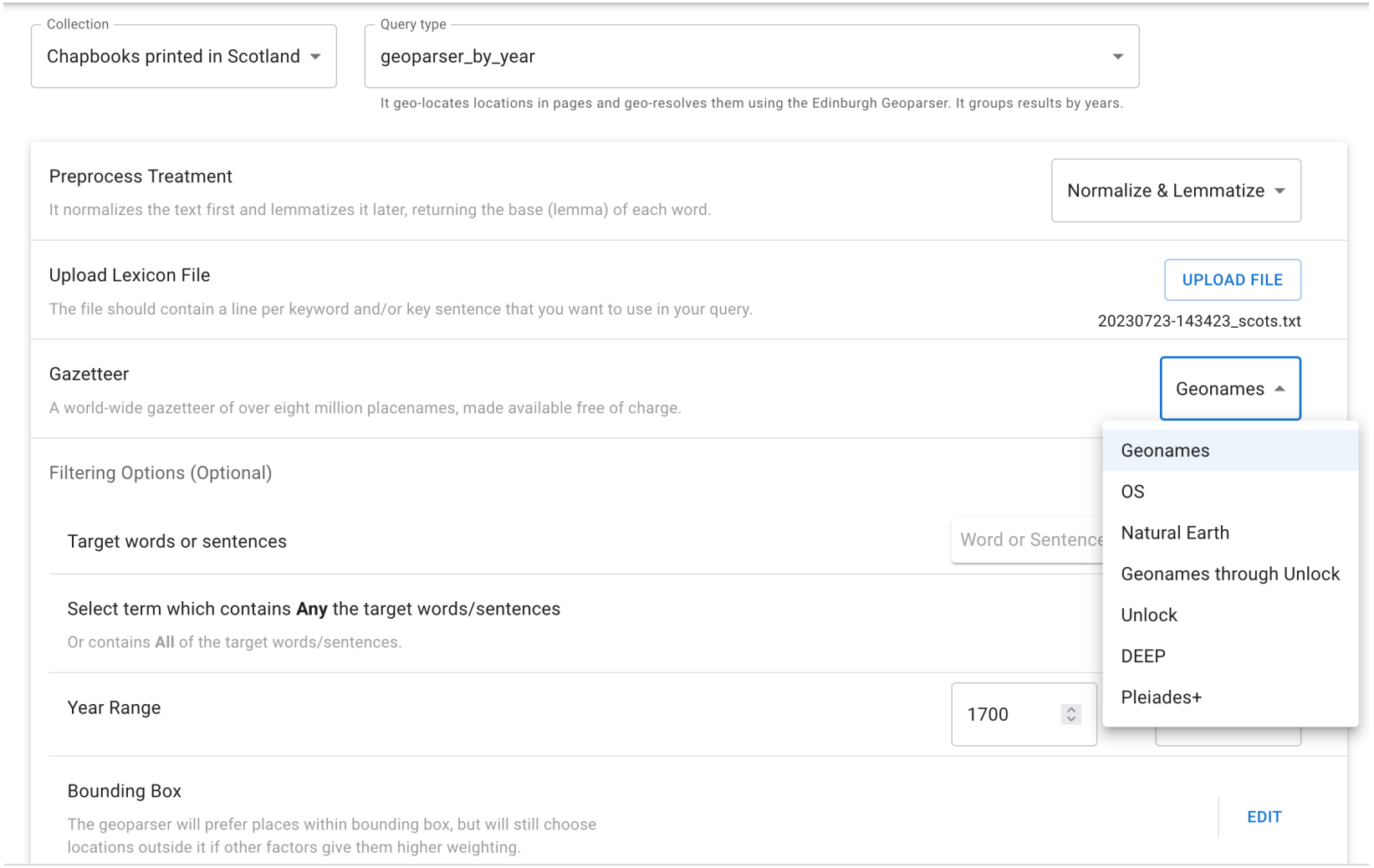
Select term which contains **Any** the target words/sentences

Or contains **All** of the target words/sentences.

defoe + Edinburgh Geoparser

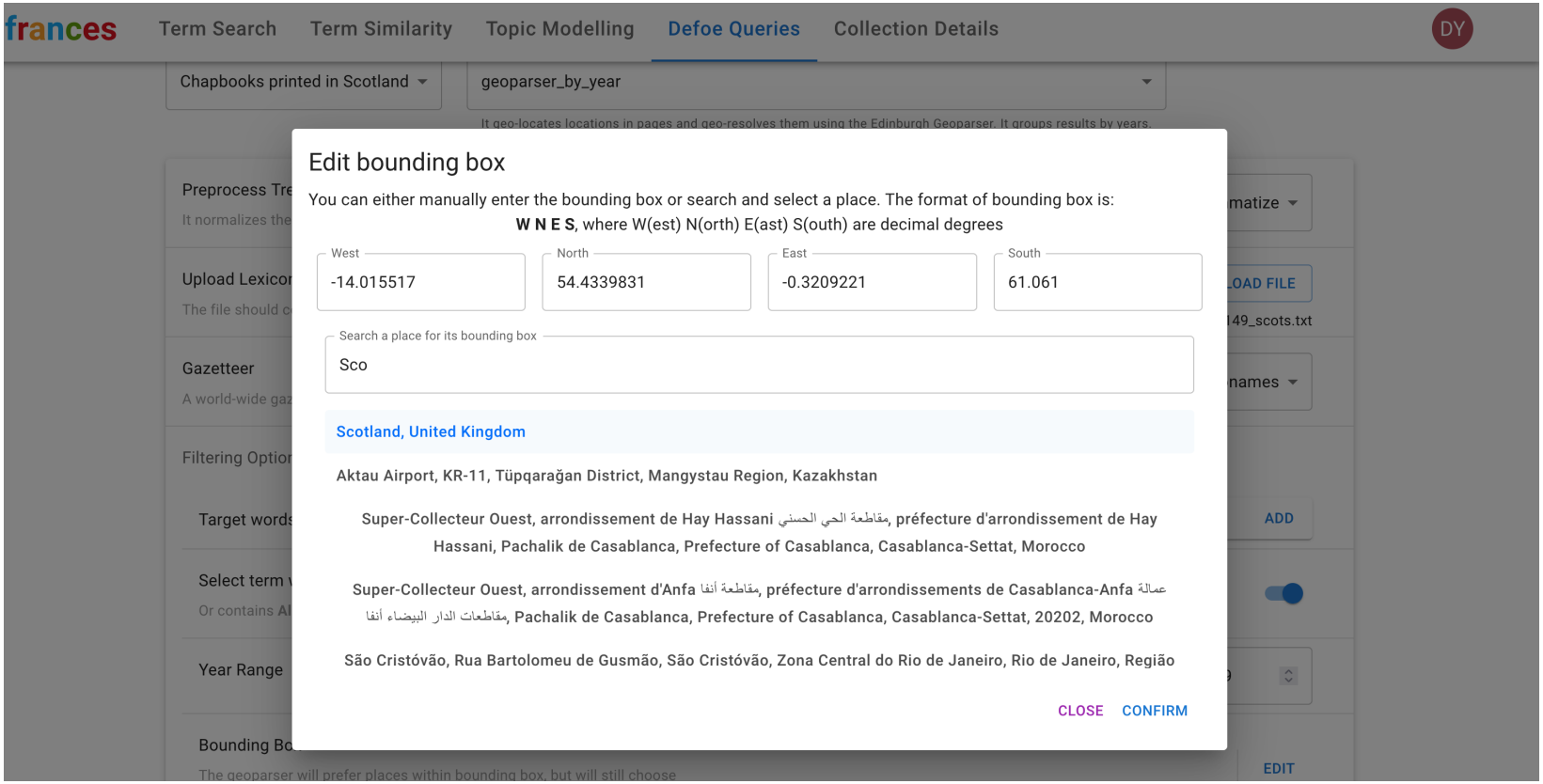
Phase 6.2: Geoparsing the Chapbooks printed in Scotland

St Andrews / CS



Phase 6.2: Geoparsing the Chapbooks printed in Scotland

St Andrews / CS



Phase 6.2: Geoparsing the Chapbooks printed in Scotland

St Andrews / CS

Defoe Queries

Collection: Chapbooks printed in Scotland

Query Type: geoparser_by_year

Lexicon Filename: scots.txt

Preprocess: Normalize & Lemmatize

Gazetteer: Geonames

Year range: 1700 - 1899

Bounding Box: -14.015517 54.4339831 -0.3209221 61.061

Submitted time: 2023-06-14 11:55:08.393946

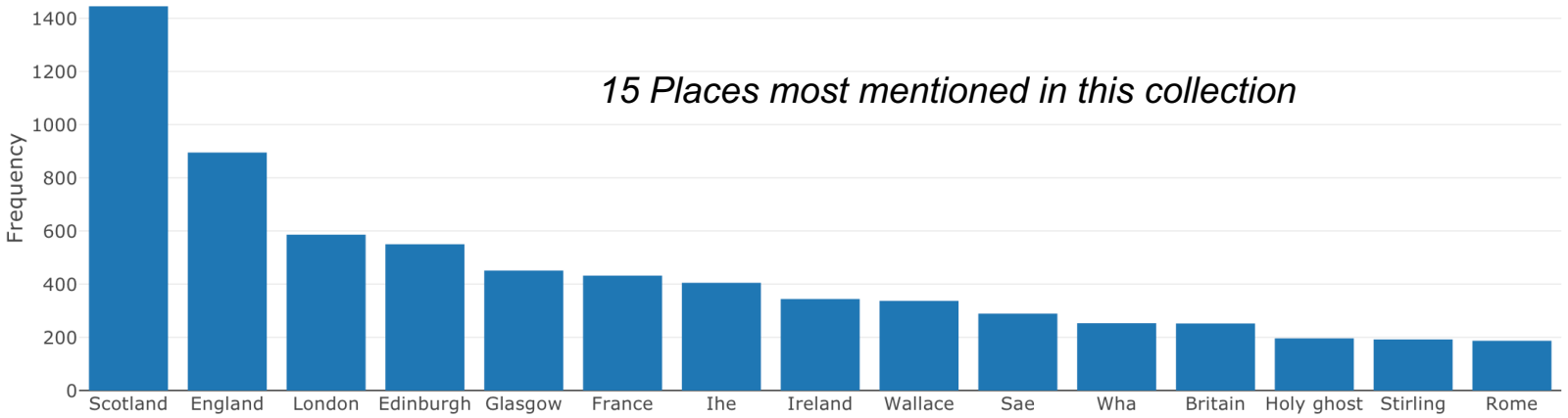
CREATE ANOTHER QUERY | CHECK ALL QUERY TASKS

Result

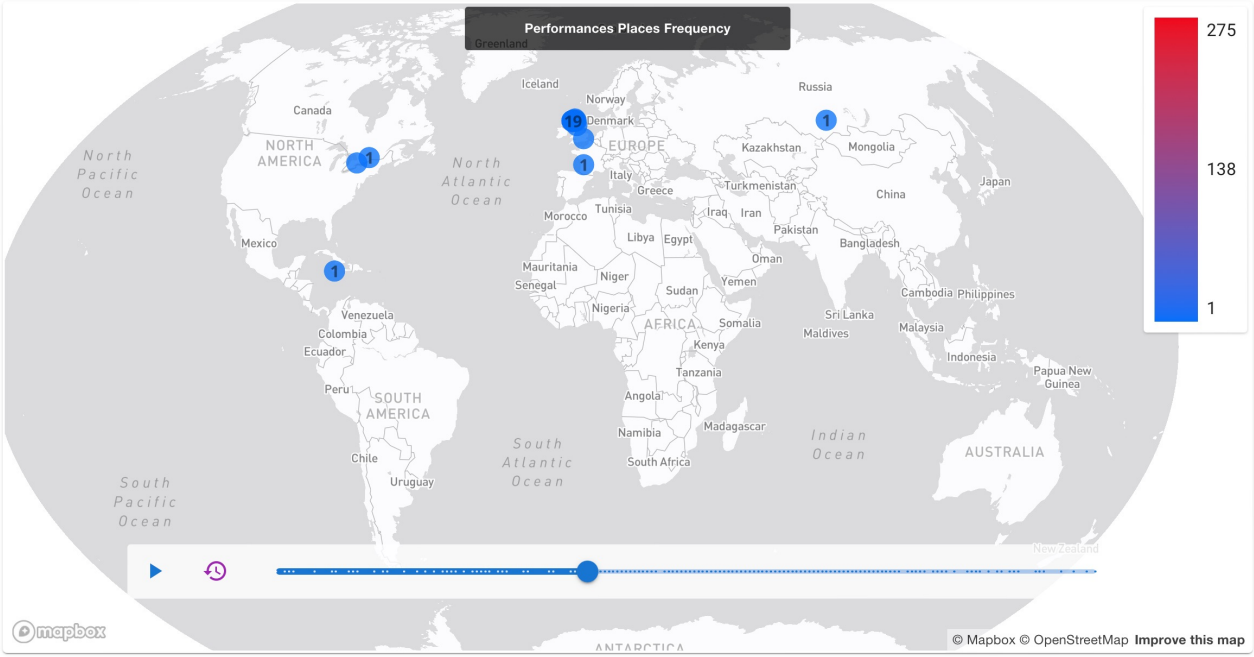
DOWNLOAD

Year	Series	Volume	Volume ID	Volume Title	Page	Words	Part	Geo
1701	204	1	104184313	unnnatural sic son	5	267	None	Spi
	204	1	104184313	unnnatural sic son	3	274	None	Tryal
	204	1	104184313	unnnatural sic son	7	276	None	Sirj
	204	1	104184313	unnnatural sic son	1	170	None	Tryal
	204	1	104184313	unnnatural sic son	6	269	None	Man the
	204	1	104184313	unnnatural sic son	2	280	None	Debaucherie CO Wi
1703	206	1	104184316	King of France His catechism	3	167	None	Europe
	206	1	104184316	King of France His catechism	6	354	None	Italy Vigo Wha

Phase 6.2: Geoparsing the Chapbooks printed in Scotland



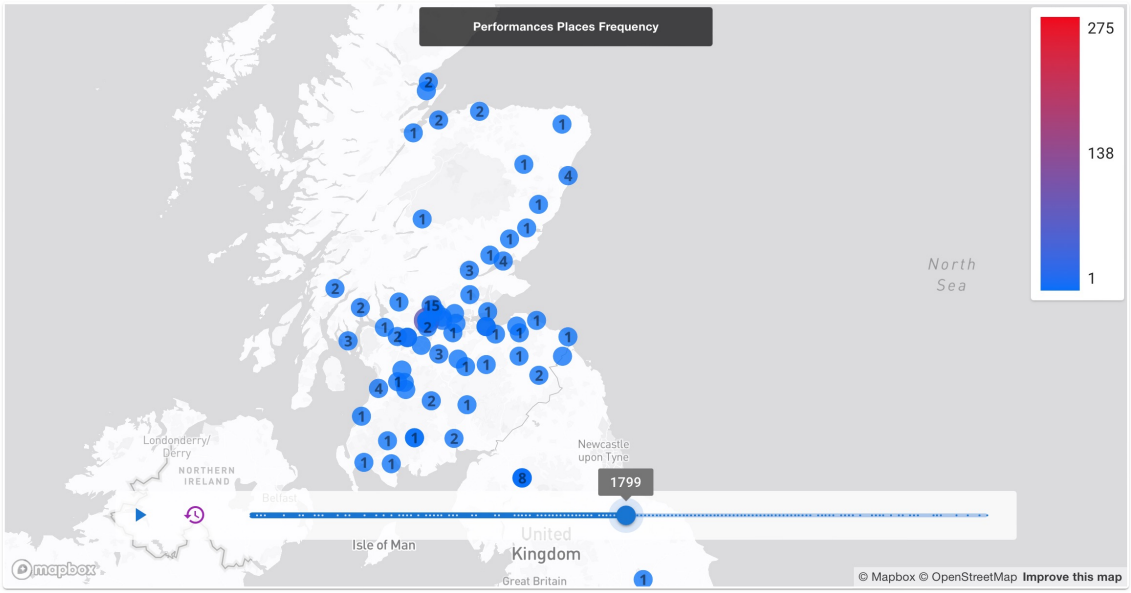
Map of geo-resolved places frequencies over time



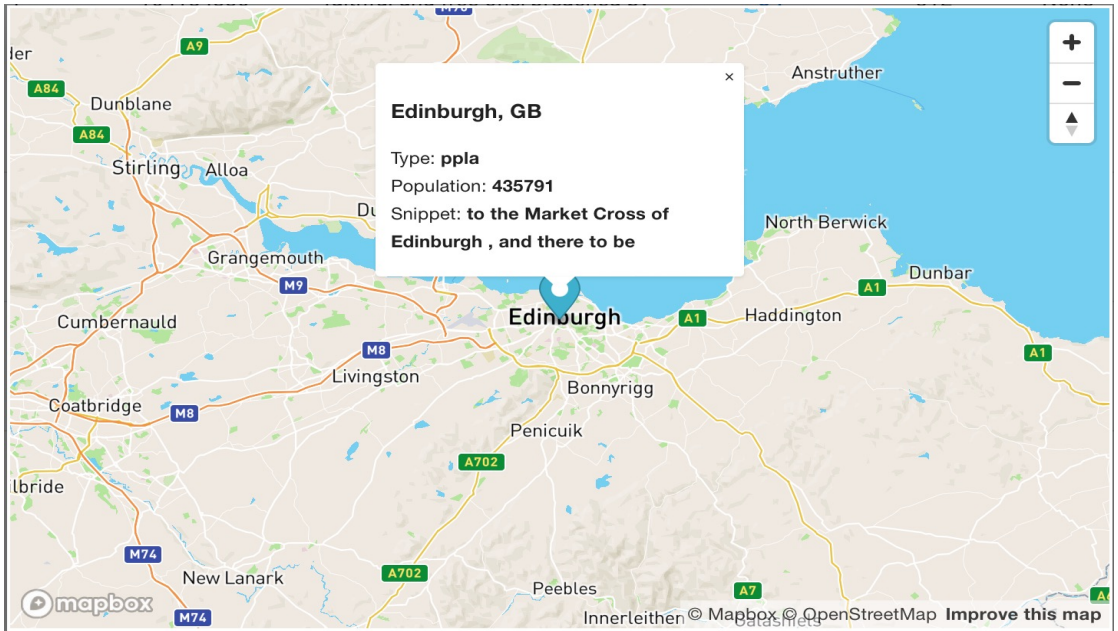
By geoparsing the chapbooks using `frances`, historians gain the ability to unravel the spatial dimensions of Scottish literature from the 18th and 19th centuries.

Phase 6.2: Geoparsing the Chapbooks printed in Scotland

St Andrews / CS



Zoom in 'Scotland'.
Frequency of geolocated places in 1799



Geolocation of "Edinburgh"

frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

St Andrews / CS

Conclusions

- *frances* enables for analysing digital historical textual collections - scalability and performance
- Cloud-based architecture ensures maintainability, efficient deployment, and scalability
- NLS and EB Ontologies standardizes representation of NLS Data Foundry collections
- Enabling knowledge graphs creation
- Deep Learning Analysis - New suite of ML functionalities that can be used to analyse textual collect.
- Parallel Text Mining queries

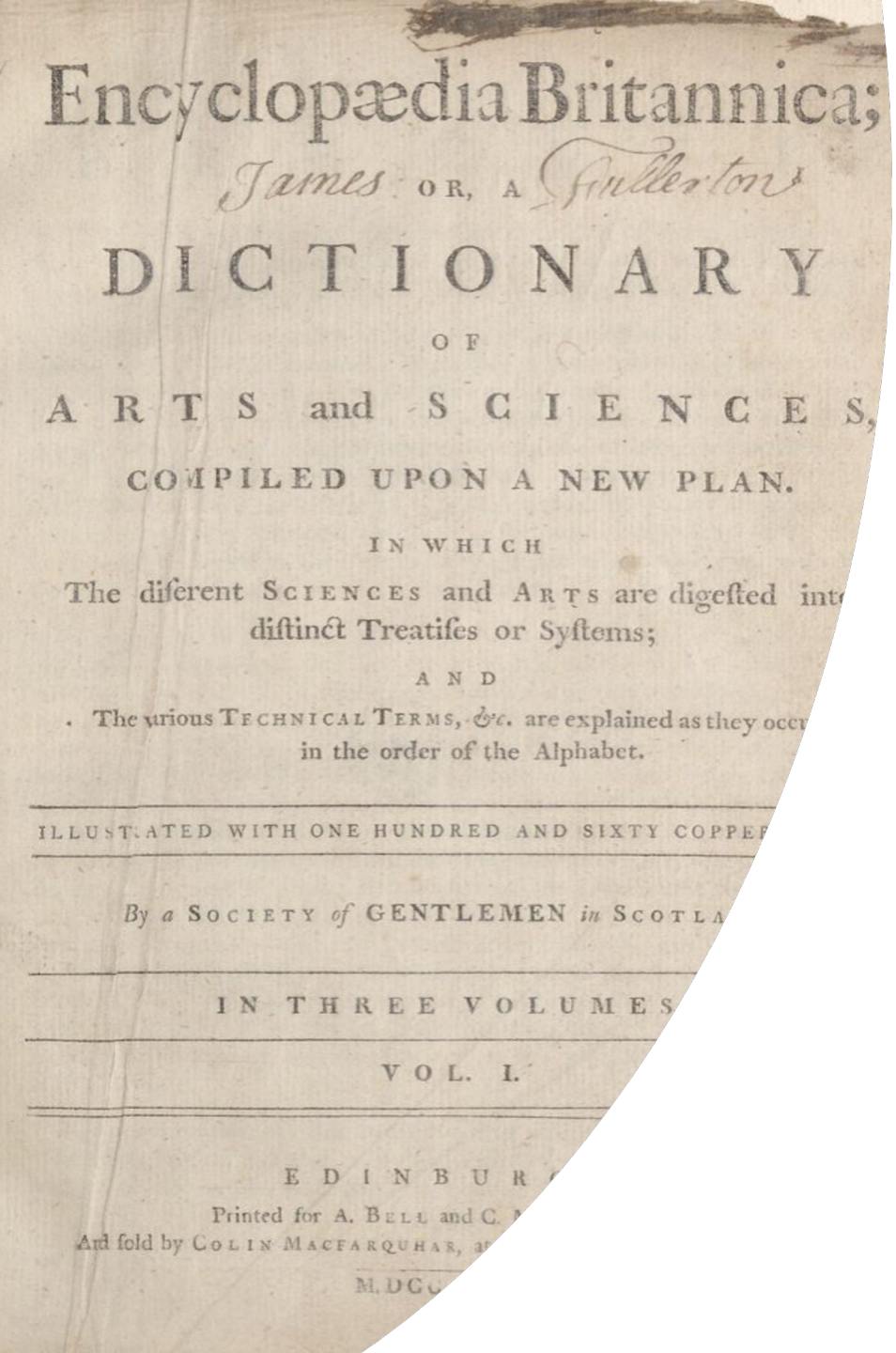
Future Work

- Expanding knowledge graph creation, collaborating with other institutions, and extensive evaluation
- Integrating external linked data sources, and gathering user feedback

Publications

frances*: a deep learning NLP and text mining web tool to unlock historical digital collections: a case study on the *Encyclopaedia Britannica Filgueira, R., 14 Dec **2022**, *2022 IEEE 18th International Conference on e-Science (e-Science)*

frances: cloud-based historical text mining with deep learning and parallel processing Yu, L., Charlton, A., Askins, W., Terras, M. & Filgueira, R., 12 Jul **2023**, (Accepted/In press) *2023 IEEE 19th International Conference on e-Science (e-Science)*


Encyclopædia Britannica;
James OR, A Fullerton's
DICTIONARY
OF
ARTS and SCIENCES,
COMPILED UPON A NEW PLAN.
IN WHICH
The different SCIENCES and ARTS are digested int
distinct Treatises or Systems;
AND
The various TECHNICAL TERMS, &c. are explained as they occ
in the order of the Alphabet.
ILLUSTRATED WITH ONE HUNDRED AND SIXTY COPPE
By a SOCIETY of GENTLEMEN in SCOTLA
IN THREE VOLUMES
VOL. I.
EDINBUR
Printed for A. BELL and C.
And sold by COLIN MACFARQUHAR, a
M.DC

Questions ?

Frances: A Deep Learning NLP and Text Mining Digital Platform for Analysis of Historical Texts at Scale

Dr. Rosa Filgueira,
Lecturer at the School of Computer Science,
University of St Andrews,
Email: rf208@st-Andrews.ac.uk