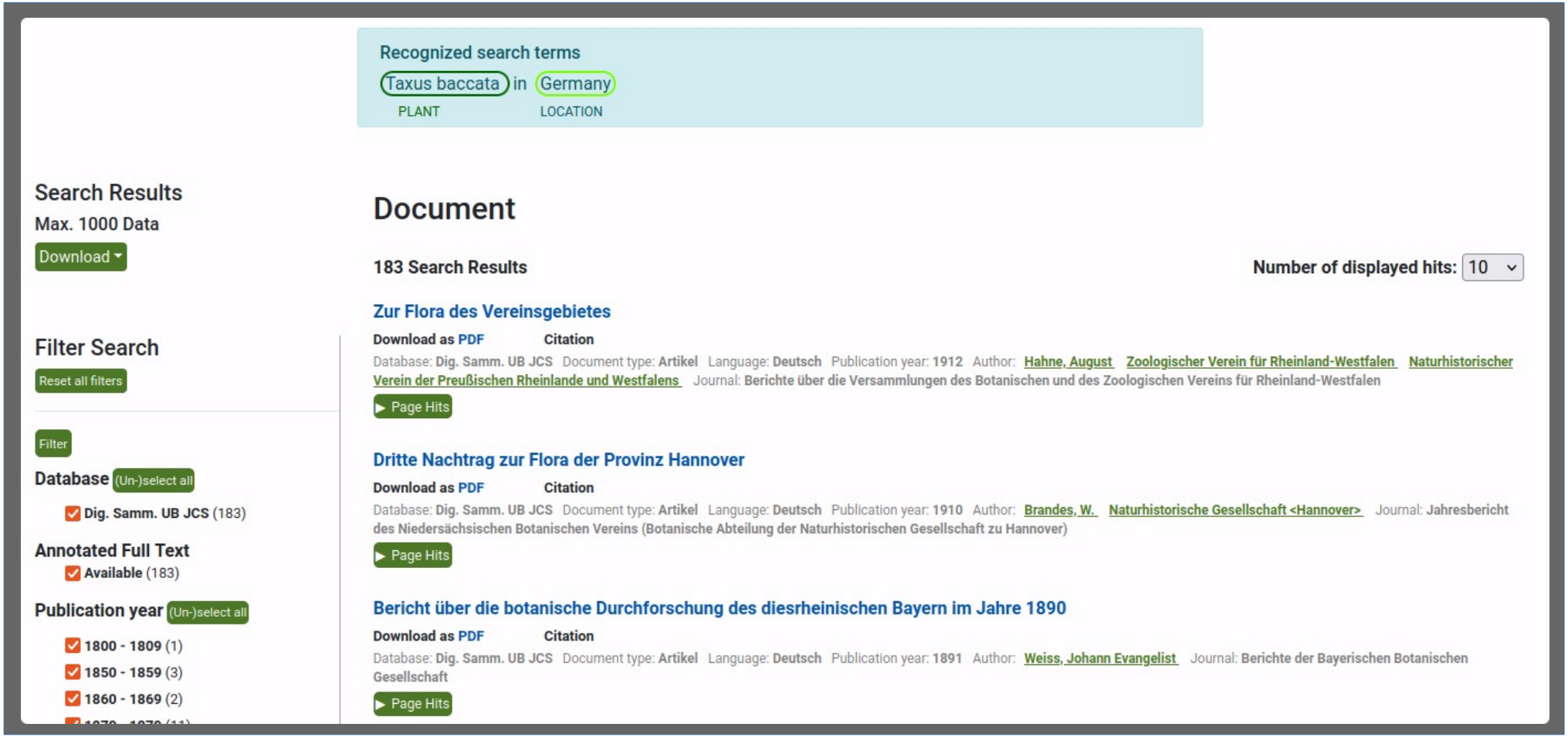
BIOfid: Accessing legacy literature the semantic (search) way

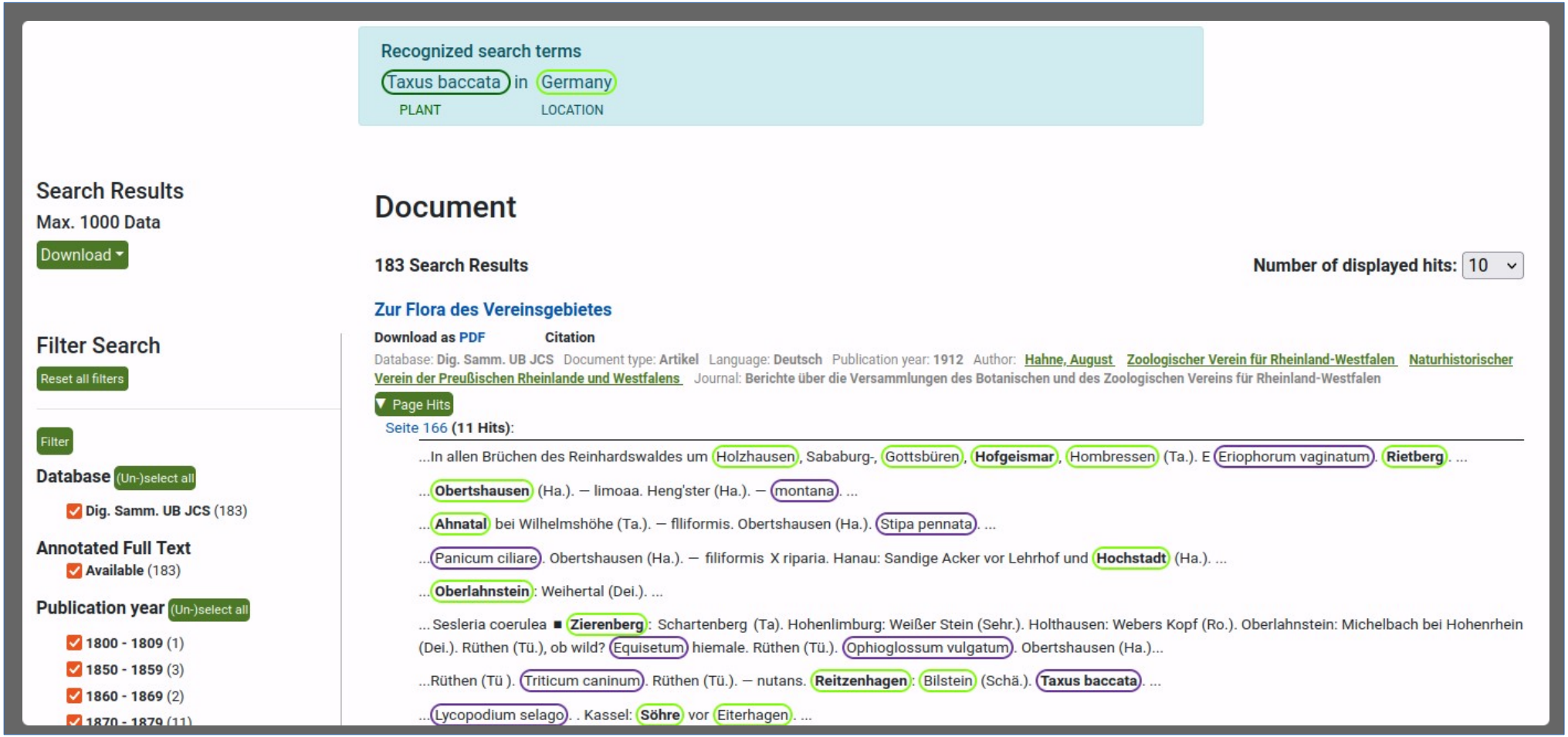
Adrian Pachzelt
Frankfurt University Library, Germany
UB Labs

The BIOfid Project

https://www.biofid.de

- Since 2017 funded by the DFG
- Text-Annotation and Text-Mining of legacy biodiversity literature (in German)
- Focus taxa: Vascular Plants, Butterflies, and Birds
- Tools
  - Text-Annotation (Demo)
  - Bio-Ontologies (Link)
  - **Semantic Search on Documents** (Link)


Recognized search terms
Taxus baccata in Germany
PLANT
LOCATION
Search Results
Max. 1000 Data
Download
Filter Search
Reset all filters
Filter
Database (Un-)select all
Dig. Samm. UB JCS (183)
Annotated Full Text
Available (183)
Publication year (Un-)select all
1800 - 1809 (1)
1850 - 1859 (3)
1860 - 1869 (2)
Document
183 Search Results
Number of displayed hits: 10
Zur Flora des Vereinsgebietes
Download as PDF
Citation
Database: Dig. Samm. UB JCS Document type: Artikel Language: Deutsch Publication year: 1912 Author: Hahne, August Zoologischer Verein für Rheinland-Westfalen Naturhistorischer Verein der Preußischen Rheinlande und Westfalens Journal: Berichte über die Versammlungen des Botanischen und des Zoologischen Vereins für Rheinland-Westfalen
Page Hits
Dritte Nachtrag zur Flora der Provinz Hannover
Download as PDF
Citation
Database: Dig. Samm. UB JCS Document type: Artikel Language: Deutsch Publication year: 1910 Author: Brandes, W. Naturhistorische Gesellschaft <Hannover> Journal: Jahresbericht des Niedersächsischen Botanischen Vereins (Botanische Abteilung der Naturhistorischen Gesellschaft zu Hannover)
Page Hits
Bericht über die botanische Durchforschung des diesrheinischen Bayern im Jahre 1890
Download as PDF
Citation
Database: Dig. Samm. UB JCS Document type: Artikel Language: Deutsch Publication year: 1891 Author: Weiss, Johann Evangelist Journal: Berichte der Bayerischen Botanischen Gesellschaft
Page Hits

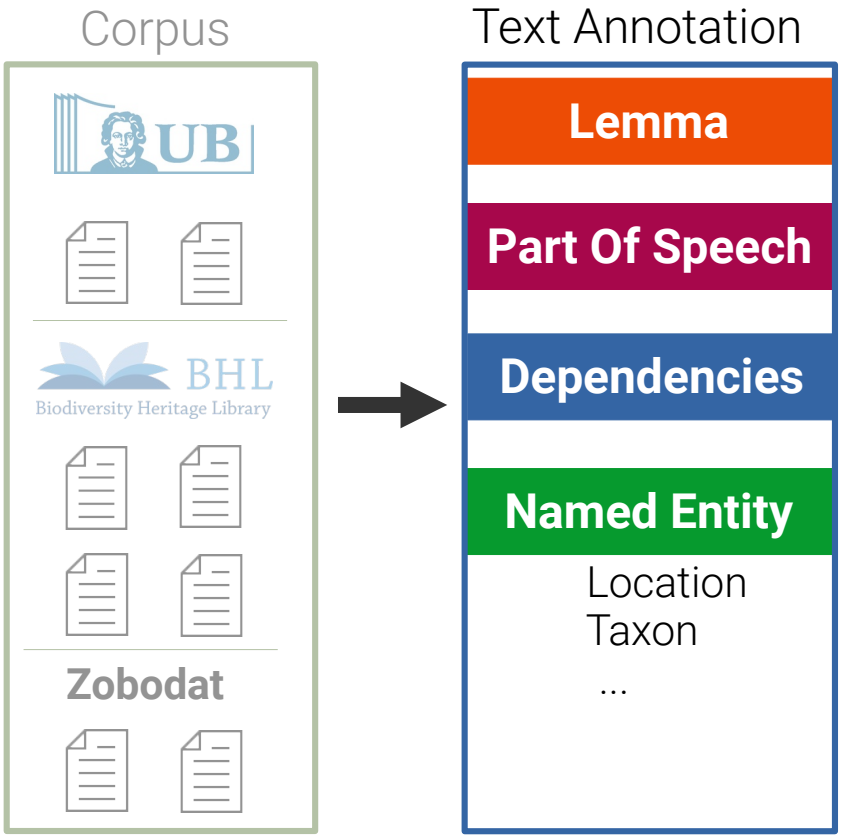

Recognized search terms
Taxus baccata in Germany
PLANT
LOCATION
Search Results
Max. 1000 Data
Download
Document
183 Search Results
Number of displayed hits: 10
Zur Flora des Vereinsgebietes
Download as PDF
Citation
Database: Dig. Samm. UB JCS Document type: Artikel Language: Deutsch Publication year: 1912 Author: Hahne, August Zoologischer Verein für Rheinland-Westfalen Naturhistorischer Verein der Preußischen Rheinlande und Westfalens Journal: Berichte über die Versammlungen des Botanischen und des Zoologischen Vereins für Rheinland-Westfalen
Page Hits
Seite 166 (11 Hits):
...In allen Brüchen des Reinhardswaldes um Holzhausen, Sababurg-, Gottsbüren, Hofgeismar, Hombressen (Ta.). E Eriophorum vaginatum. Rietberg. ...
...Obertshausen (Ha.). — limoaa. Heng'ster (Ha.). — montana. ...
...Ahnatal bei Wilhelmshöhe (Ta.). — filliformis. Obertshausen (Ha.). Stipa pennata. ...
...Panicum ciliare. Obertshausen (Ha.). — filiformis X riparia. Hanau: Sandige Acker vor Lehrhof und Hochstadt (Ha.). ...
...Oberlahnstein: Weihertal (Dei.). ...
... Sesleria coerulea ■ Zierenberg: Schartenberg (Ta). Hohenlimburg: Weißer Stein (Sehr.). Holthausen: Webers Kopf (Ro.). Oberlahnstein: Michelbach bei Hohenrhein (Dei.). Rüthen (Tü.), ob wild? Equisetum hiemale. Rüthen (Tü.). Ophioglossum vulgatum. Obertshausen (Ha.)...
...Rüthen (Tü). Triticum caninum. Rüthen (Tü.). — nutans. Reitzenhagen: Bilstein (Schä.). Taxus baccata. ...
...Lycopodium selago. . Kassel: Söhre vor Eiterhagen. ...
Filter Search
Reset all filters
Filter
Database (Un-)select all
Dig. Samm. UB JCS (183)
Annotated Full Text
Available (183)
Publication year (Un-)select all
1800 - 1809 (1)
1850 - 1859 (3)
1860 - 1869 (2)
1870 - 1879 (11)

Forging Data

BIOfid Annotation Workflow

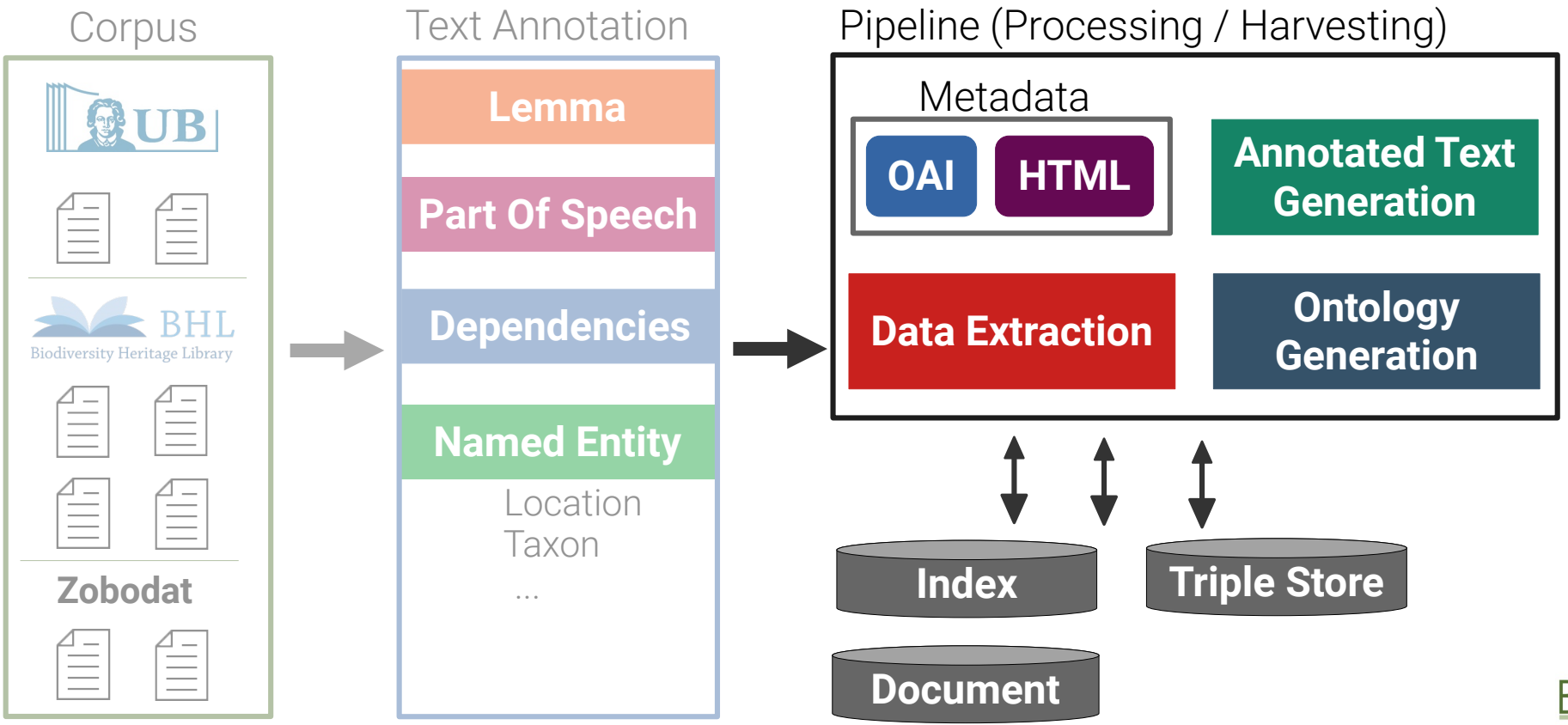
File Icon by Dario Ferrando

BIOfid Annotation Workflow



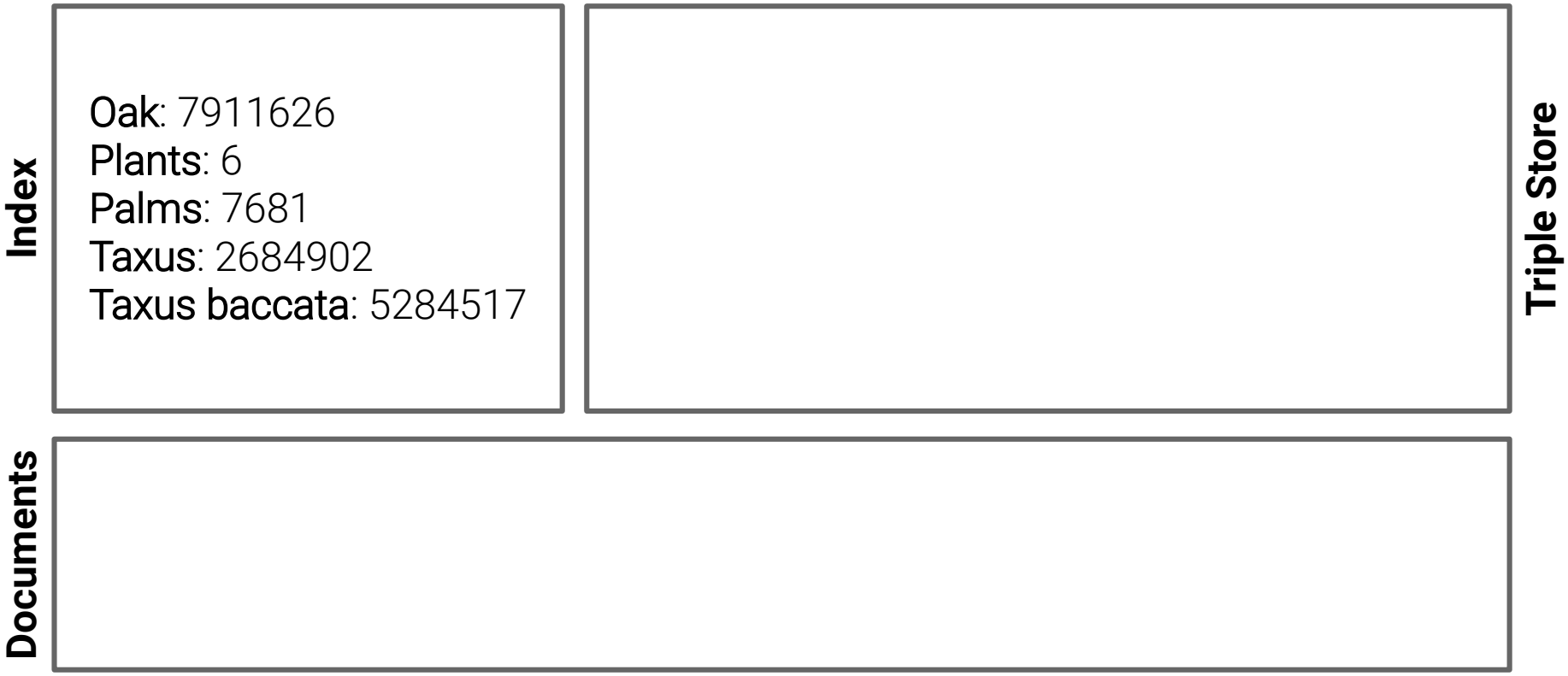
File Icon by Dario Ferrando

BIOfid Annotation Workflow



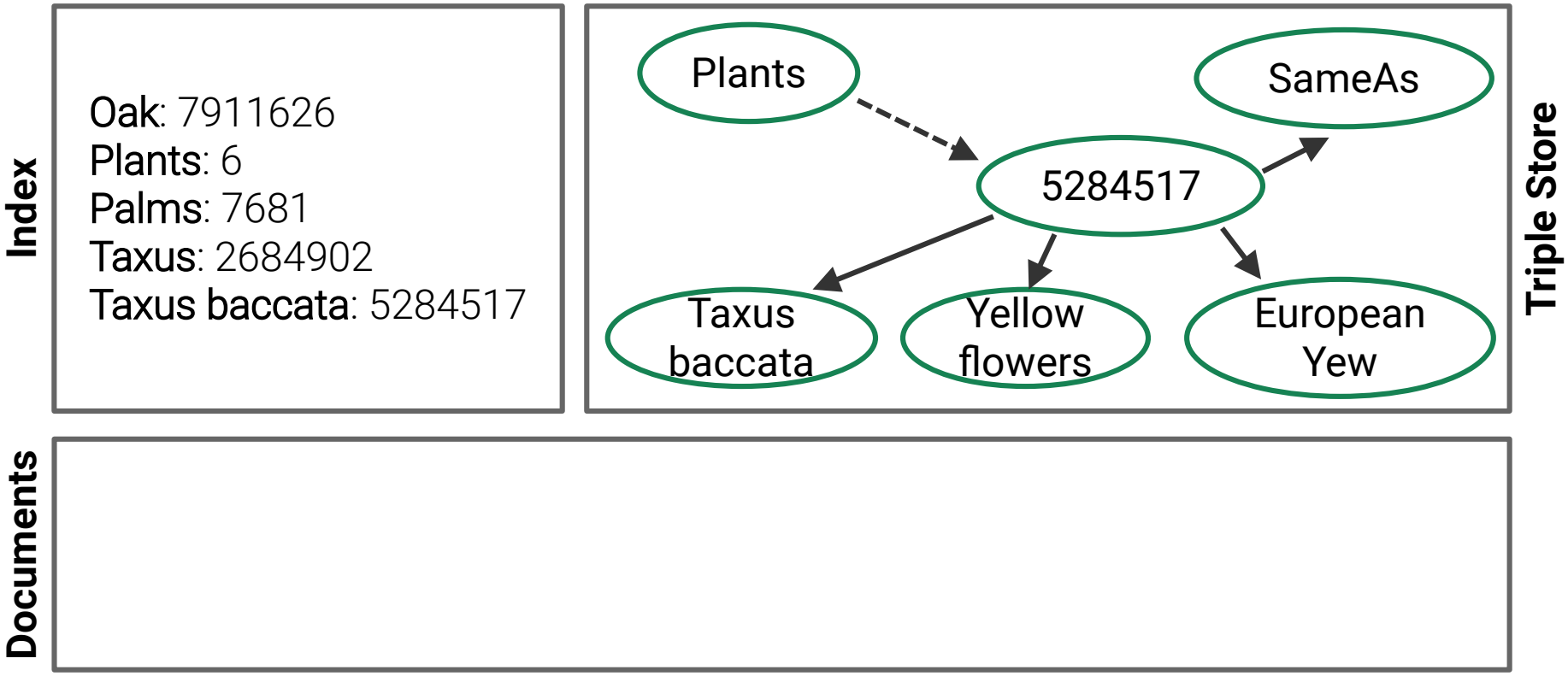
File Icon by Dario Ferrando

Linking the Data



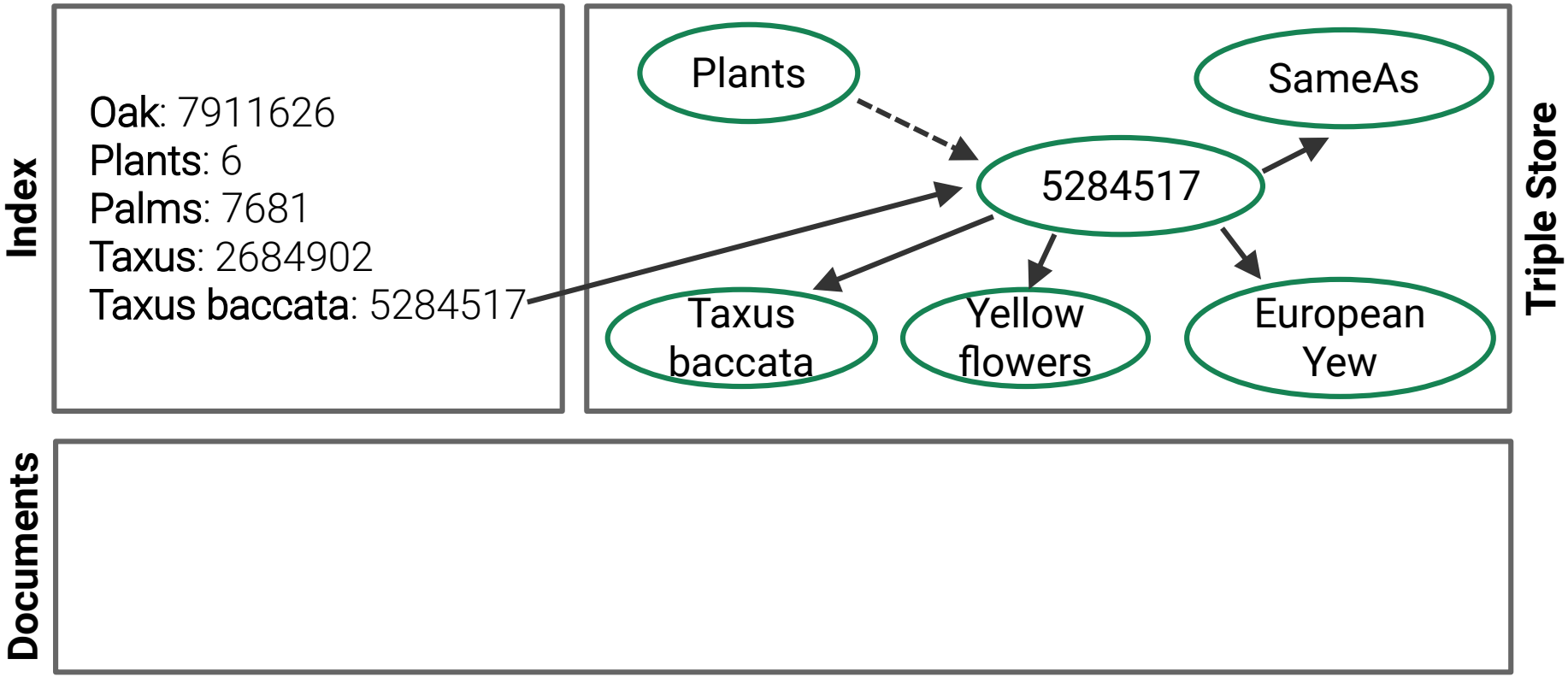
Spoiler: The IDs are URIs, but let's make this simple.

Linking the Data



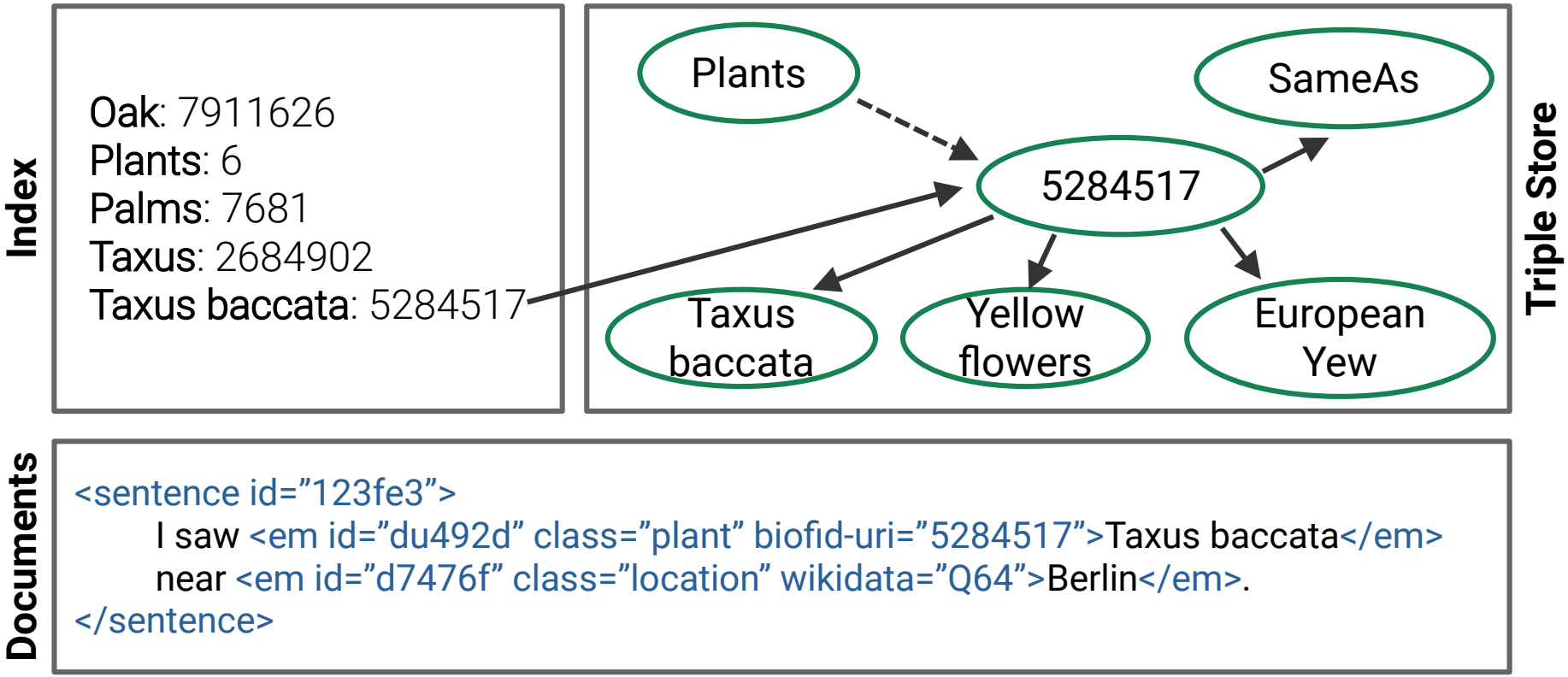
Spoiler: The IDs are URIs, but let's make this simple.

Linking the Data



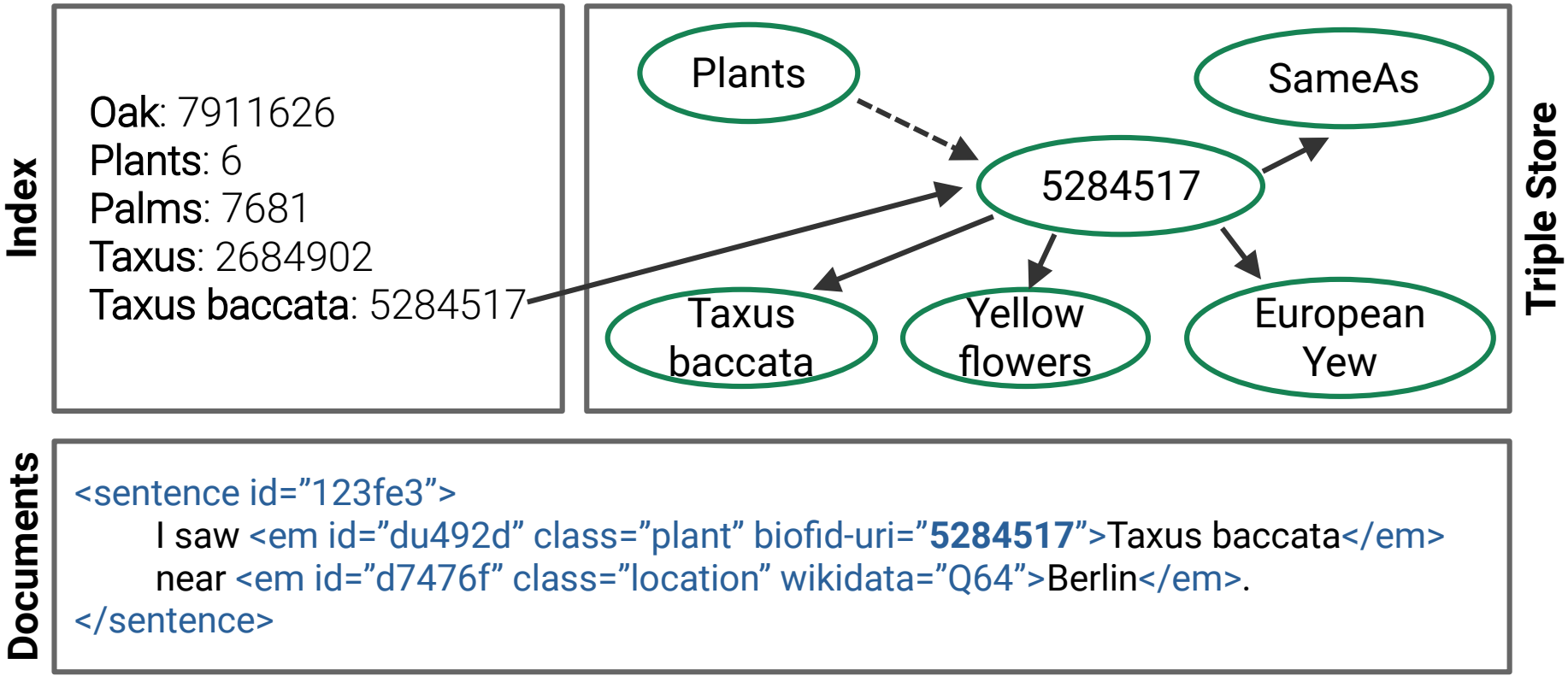
Spoiler: The IDs are URIs, but let's make this simple.

Linking the Data



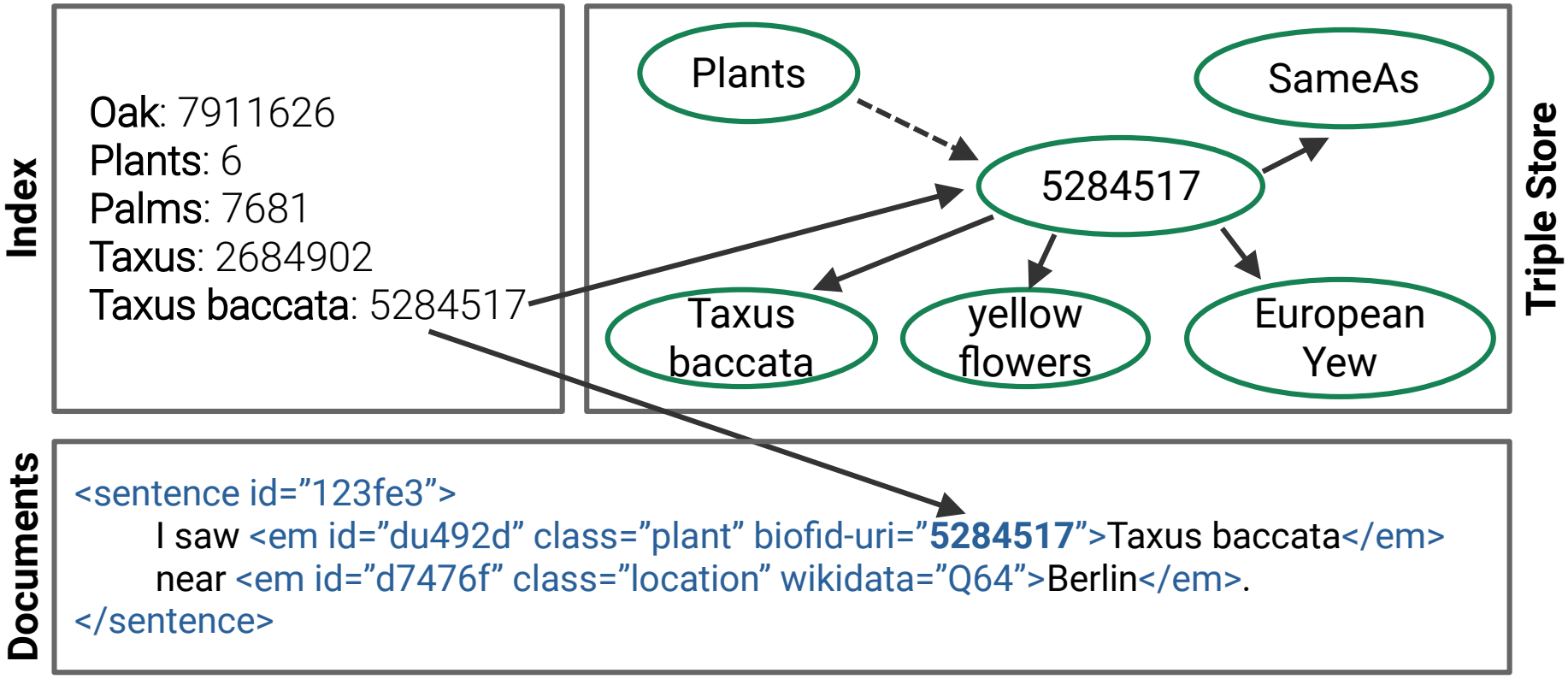
Spoiler: The IDs are URIs, but let's make this simple.

Linking the Data



Spoiler: The IDs are URIs, but let's make this simple.

Linking the Data



Spoiler: The IDs are URIs, but let's make this simple.

Getting the Data out

The BIOfid Semantic Search Engine

Query Analysis

Plants with yellow flowers

The BIOfid Semantic Search Engine

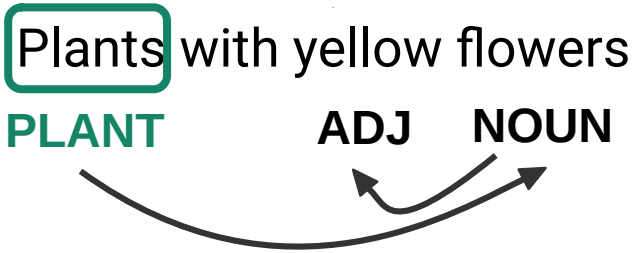
Query Analysis

Plants with yellow flowers **Natural Language Processing**

PLANT **ADJ** **NOUN**

The BIOfid Semantic Search Engine

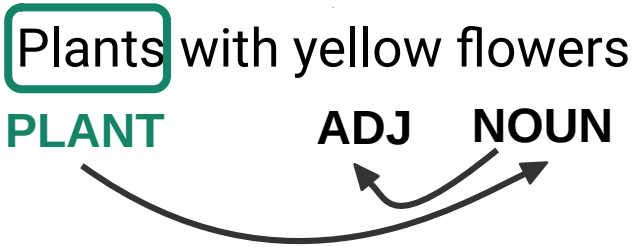
Query Analysis



Natural Language Processing

The BIOfid Semantic Search Engine

Query Analysis

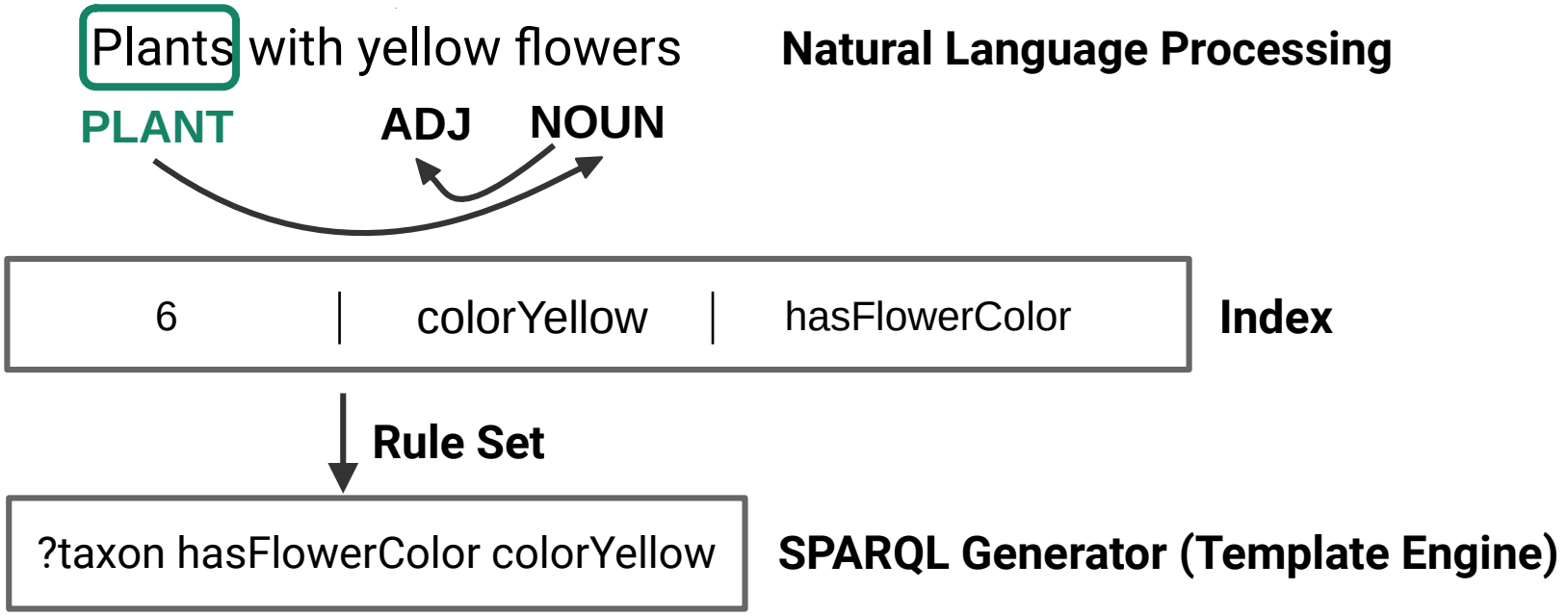


Natural Language Processing

Index

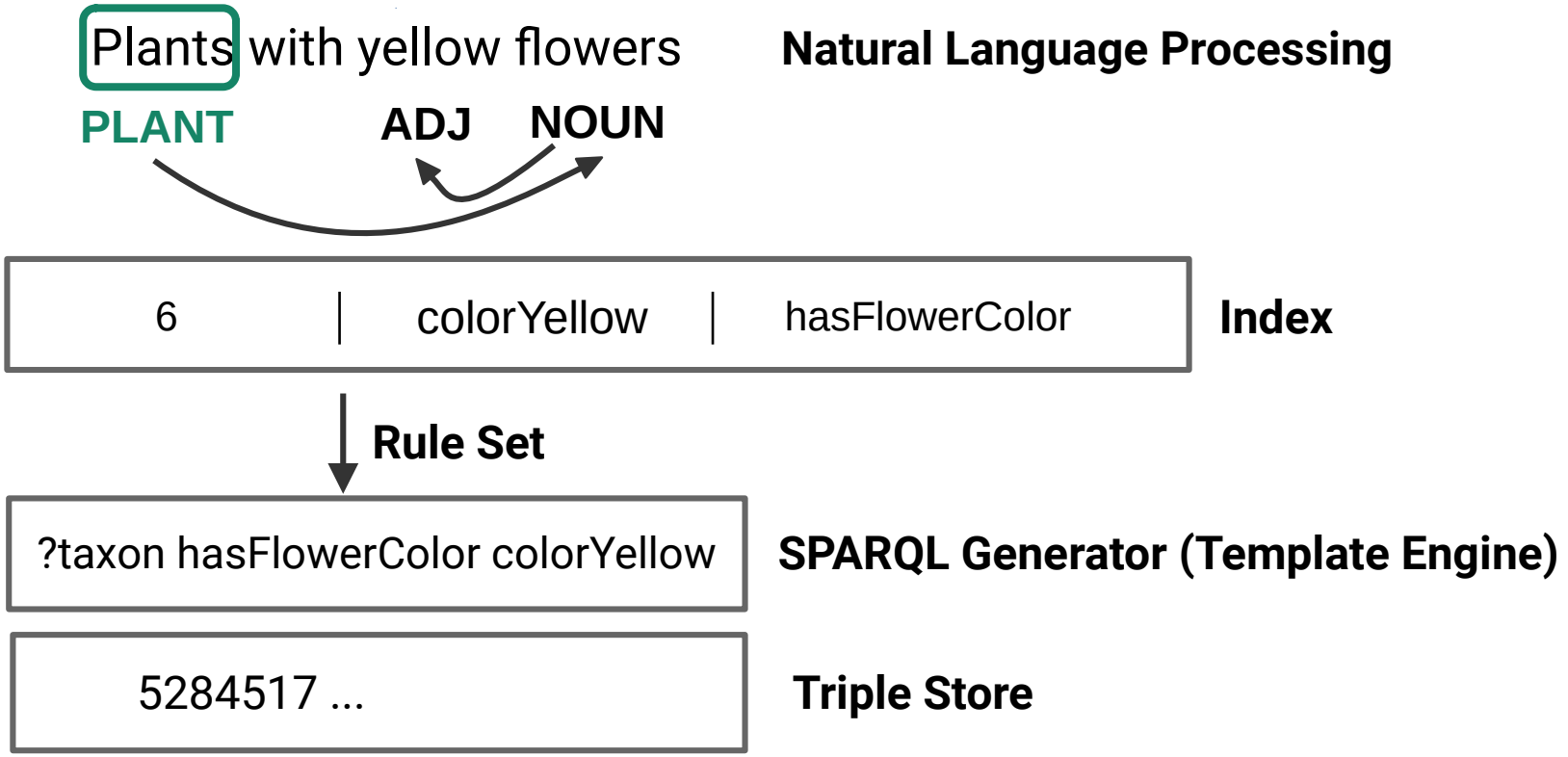
The BIOfid Semantic Search Engine

Query Analysis



The BIOfid Semantic Search Engine

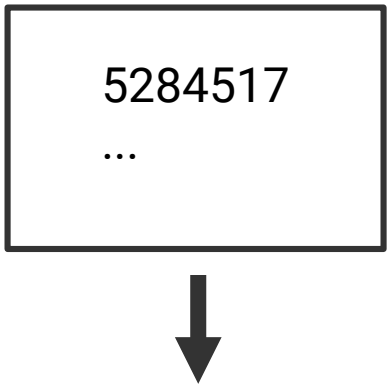
Query Analysis



The BIOfid Semantic Search Engine

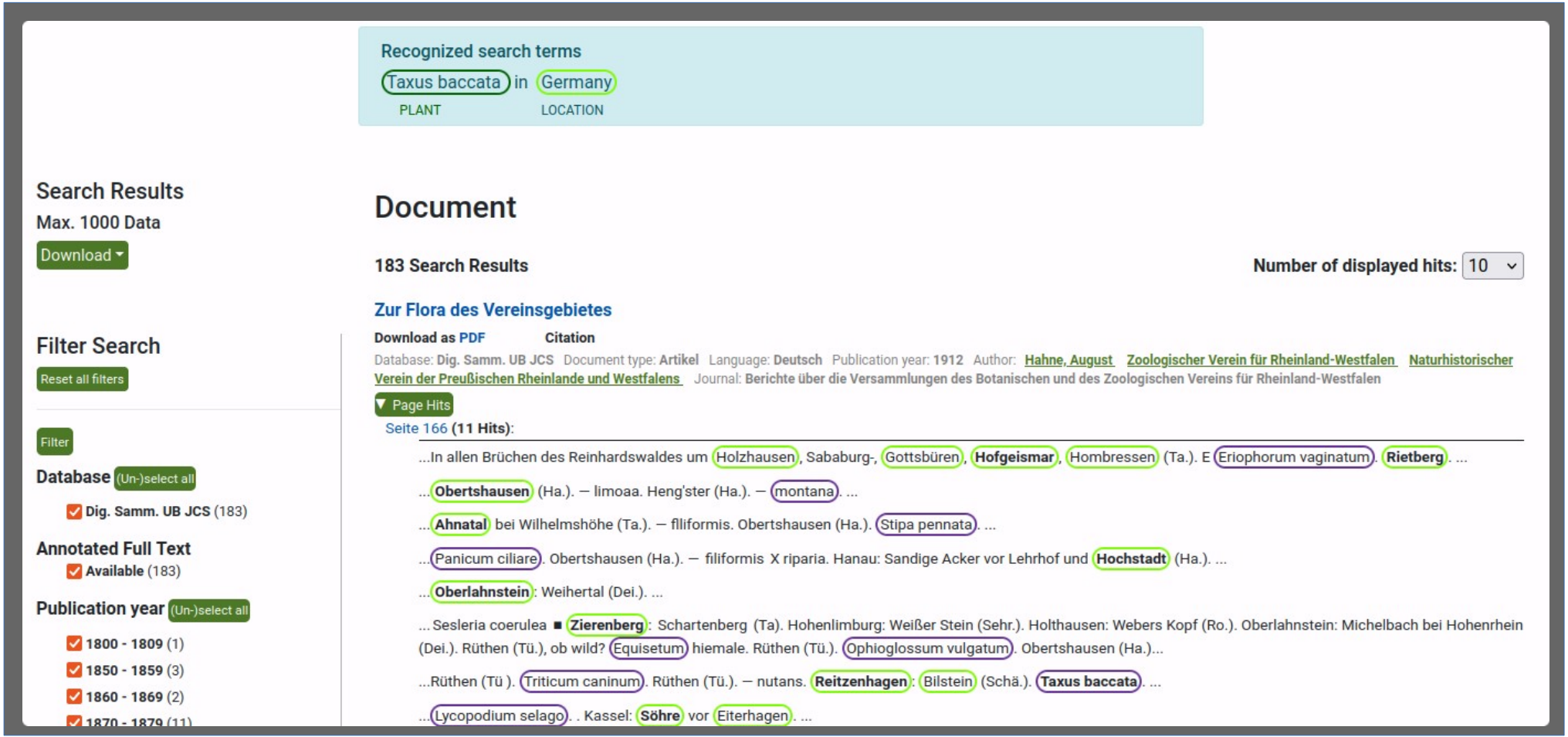
Getting the Documents

Triple Store Response



Documents

```
<sentence id="123fe3">
    I saw <em id="du492d" class="plant" biofid-uri="5284517">Taxus baccata</em>
    near <em id="d7476f" class="location" wikidata="Q64">Berlin</em>.
</sentence>
```


Recognized search terms
Taxus baccata in Germany
PLANT
LOCATION
Search Results
Max. 1000 Data
Download
Document
183 Search Results
Number of displayed hits: 10
Zur Flora des Vereinsgebietes
Download as PDF
Citation
Database: Dig. Samm. UB JCS Document type: Artikel Language: Deutsch Publication year: 1912 Author: Hahne, August Zoologischer Verein für Rheinland-Westfalen Naturhistorischer Verein der Preußischen Rheinlande und Westfalens Journal: Berichte über die Versammlungen des Botanischen und des Zoologischen Vereins für Rheinland-Westfalen
Page Hits
Seite 166 (11 Hits):
...In allen Brüchen des Reinhardswaldes um Holzhausen, Sababurg-, Gottsbüren, Hofgeismar, Hombressen (Ta.). E Eriophorum vaginatum. Rietberg. ...
...Obertshausen (Ha.). — limoaa. Heng'ster (Ha.). — montana. ...
...Ahnatal bei Wilhelmshöhe (Ta.). — filliformis. Obertshausen (Ha.). Stipa pennata. ...
...Panicum ciliare. Obertshausen (Ha.). — filiformis X riparia. Hanau: Sandige Acker vor Lehrhof und Hochstadt (Ha.). ...
...Oberlahnstein: Weihertal (Dei.). ...
... Sesleria coerulea ■ Zierenberg: Schartenberg (Ta). Hohenlimburg: Weißer Stein (Sehr.). Holthausen: Webers Kopf (Ro.). Oberlahnstein: Michelbach bei Hohenrhein (Dei.). Rüthen (Tü.), ob wild? Equisetum hiemale. Rüthen (Tü.). Ophioglossum vulgatum. Obertshausen (Ha.)...
...Rüthen (Tü). Triticum caninum. Rüthen (Tü.). — nutans. Reitzenhagen: Bilstein (Schä.). Taxus baccata. ...
...Lycopodium selago. . Kassel: Söhre vor Eiterhagen. ...
Filter Search
Reset all filters
Filter
Database (Un-)select all
Dig. Samm. UB JCS (183)
Annotated Full Text
Available (183)
Publication year (Un-)select all
1800 - 1809 (1)
1850 - 1859 (3)
1860 - 1869 (2)
1870 - 1879 (11)

Lessons learned

- Store each NLP enrichment state of the user query

  You may need the data in some later (code) expansion

- Prefer templates for query interpretation over rules

  The query space is smaller than you think! Example: Lango

- Get ahead of drowning in data

  Can the application handle large (!) amounts of database responses?

- Do not underestimate the power of normalized data

  It creates a semantic context in your document database out-of-the-box

Alternatives

- Topic Modelling

  Latent Semantic Analysis, Latent Dirichlet Allocation, …
  May not scale well
  Some algorithms may be hard to parametrize

- Machine Learning

  Getting the training data is hard.
  Re-training every time the ontological knowledge changes (?)

Disclaimer: I applied neither of these approaches myself!

Information Extraction

Work in Progress

```
<sentence id="123fe3">
    I saw <em id="du492d" class="plant" biofid-uri="5284517">Taxus baccata</em>
    near <em id="d7476f" class="location" wikidata="Q64">Berlin</em>.
</sentence>
```

```
<SimpleDarwinRecord>
    <dwc:taxonID rdf:resource="5284517"/>
    <dwc:locationID rdf:resource="https://www.wikidata.org/Q64"/>
    <dwc:associatedReferences rdf:resource="doi:10.1234/789"/>
</SimpleDarwinRecord>
```

#Nanopublications

Thank you!

- **UB Labs Blog:** https://labs.ub.uni-frankfurt.de
- **BIOfid GitHub Repos:**
  - General: https://github.com/FID-Biodiversity
  - Harvesting large document facilities: LiteratureCrawler
  - Indexing annotated texts to Solr: TaggedTextTokenizer

Appendix

The BIOfid Semantic Search Engine

Query Analysis (Alternative)

Taxus baccata in Germany

The BIOfid Semantic Search Engine

Query Analysis (Alternative)

Taxus baccata in Germany

PLANT **LOCATION**

Natural Language Processing

The BIOfid Semantic Search Engine

Query Analysis (Alternative)