

SPNHC 2022, Edinburgh 5th-10th June 2022

Workshop Title:

Application of BIOfid tools for extracting data from biodiversity literature

Length of Workshop

Half Day (9th June 2022)

Institutions:

- (1) University Library Johann Christian Senckenberg, Frankfurt am Main, Germany
- (2) Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany
- (3) Department of Computer Science and Mathematics, Goethe University Frankfurt, Frankfurt am Main, Germany

Workshop Facilitators:

- Dr. Gerwin Kasperek (1),
- Giuseppe Abrami (3)
- Dr. Christine Driller (2)
- Dr. Andy Lücking (3)
- Prof. Dr. Alexander Mehler (3)
- Carlos Alberto Martínez-Muñoz (2)
- Dr. Adrian Pachzelt (1)

Content:

In an ideal world, extraction of machine-readable data and knowledge from natural-language biodiversity literature would be done automatically, but not so currently. The BIOfid project has developed some tools that can help with important parts of this highly demanding task, while certain parts of the workflow cannot be automated yet. BIOfid focuses on the 20th century legacy literature, a large part of which is only available in printed form. In this workshop, we will present the current state of the art in mobilisation of data from our corpus, as well as some challenges ahead of us. Together with the participants, we will exercise or explain the following tasks (some of which can be performed by the participants themselves, while other tasks currently require execution by our specialists with special equipment): Preparation of text files as an input; pre-processing with TextImager/TextAnnotator; semi-automated annotation and linking of named entities; generation of output in various formats; evaluation of the output. The workshop will also provide an outlook for further developments regarding extraction of statements from natural-language literature, with the long-term aim to produce machine-readable data from literature that can extend biodiversity databases and knowledge graphs.

Main Message:

The BIOfid project has developed tools that can help with important parts of the workflows necessary to mobilise data from biodiversity literature. In this workshop, we will demonstrate the current state of the art and some challenges ahead of us. Together with the participants, we will exercise or explain several tasks necessary to mobilise this kind of data.

Audience:

The workshop is aimed at researchers in the field of biodiversity who want to mobilize data from the literature for IT-based analysis, as well as developers working on similar issues.

Application of BIOfid tools for extracting data from biodiversity literature

BIOfid Team



University Library Johann Christian Senckenberg, Frankfurt am Main



Text Technology Lab, Goethe University, Frankfurt am Main



Senckenberg Society for Nature Research, Frankfurt am Main

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline (Hands-on)
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

Three partners behind BIOfid

University Library Johann Christian Senckenberg



- comprehensive collection of biodiversity literature (printed & digital)
- service provider for nation-wide supply of literature

Senckenberg Society for Nature Research



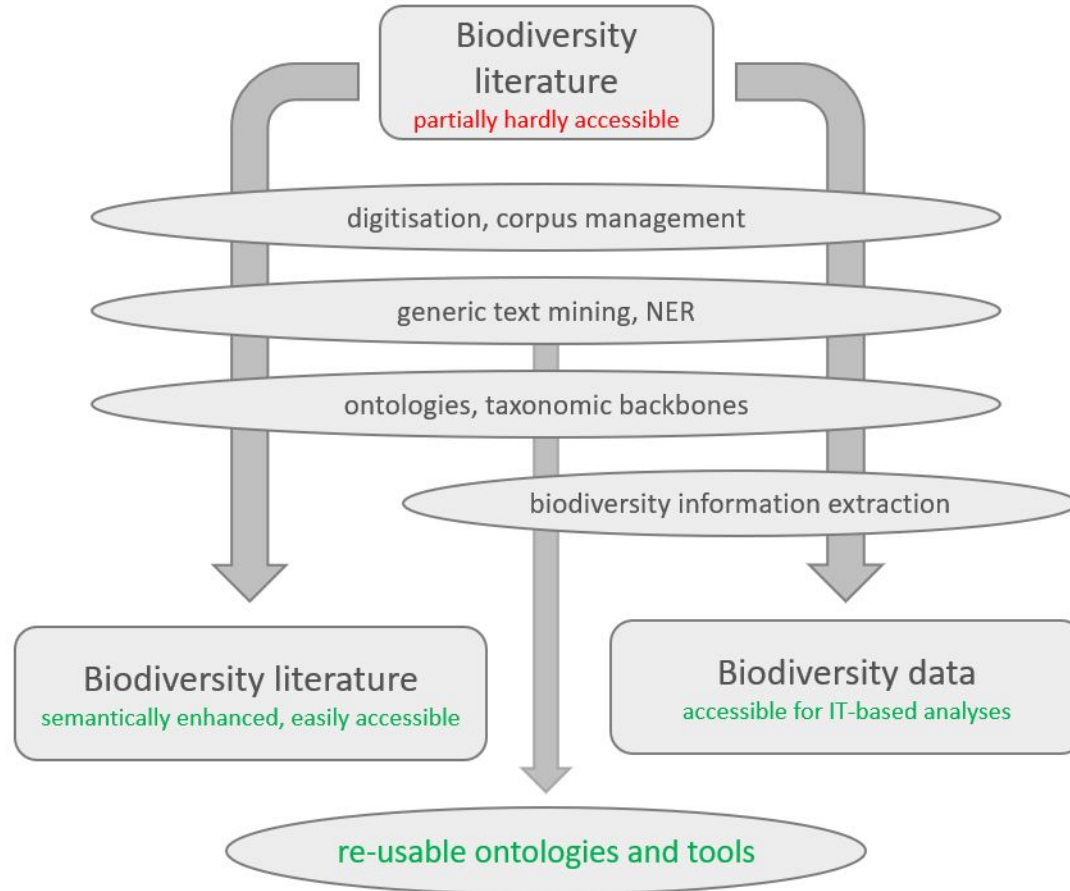
- one of the leading German institutions in biodiversity research
- expertise in biodiversity informatics

Text Technology Lab, Goethe University, Frankfurt am Main

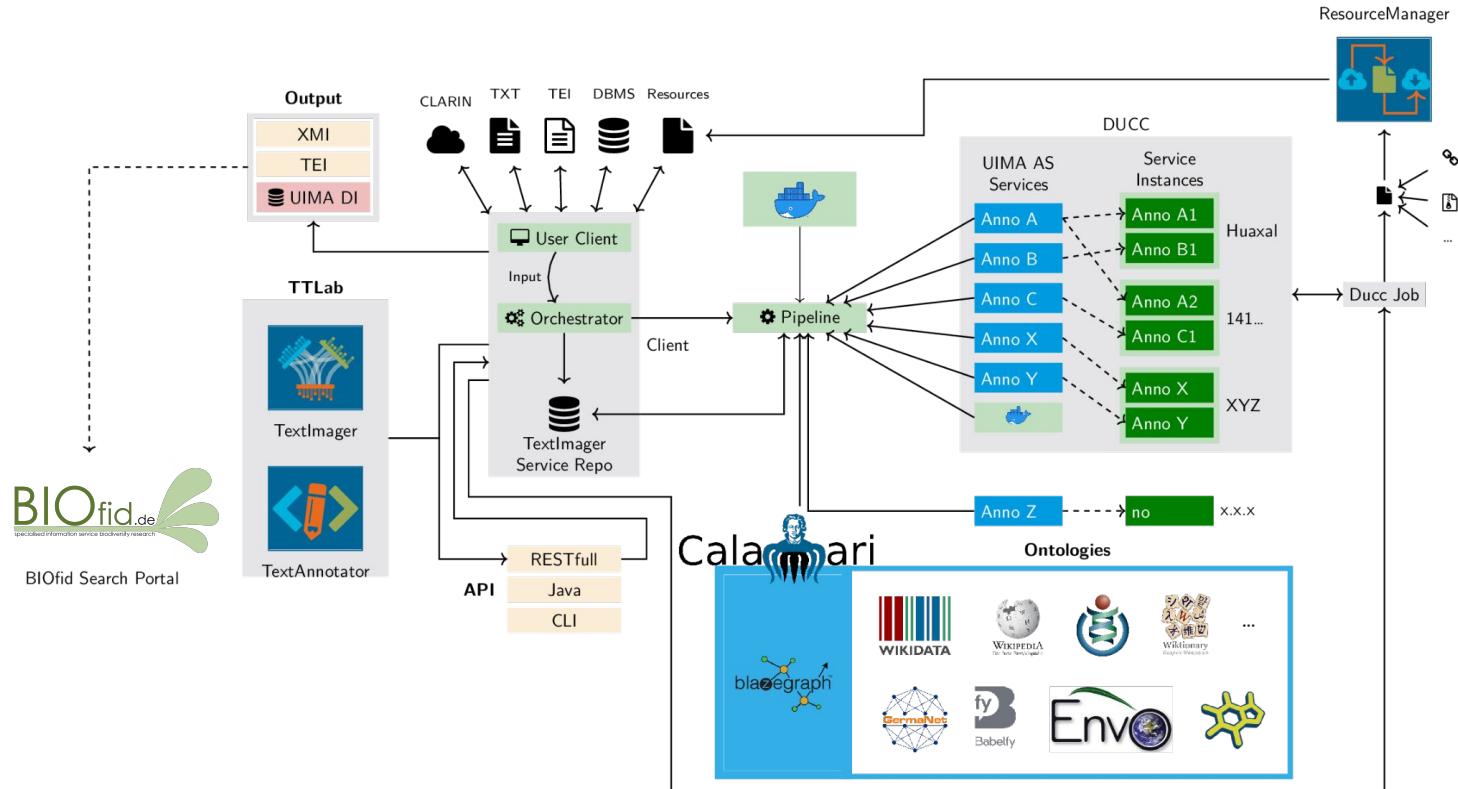


- excellent team for automated analysis and synthesis of structures and content of natural-language texts

How to make biodiversity literature more accessible



Procedures and data flows, somewhat more detailed...



Central European focus

BIOfid has a Central European focus - in particular German-speaking regions.

BIOfid (phase 1 & 2) is focussed on certain taxonomic groups - mainly higher plants, birds, butterflies, but also myriapods, insects...

But, in the end, BIOfid will provide services relevant to other regions and/or taxonomic groups as well.



Digitisation of printed literature

UB JCS has largest collection of print biology literature in Central Europe... and strives to make large parts available on-line

- “German Botanical Journals, 1753–1914”
project completed 2013, ca. 180 titles
<https://sammlungen.ub.uni-frankfurt.de/botanik?lang=en>
- 20th century biodiversity literature from Central Europe
ongoing part of BIOfid project; in-copyright-materials
<https://sammlungen.ub.uni-frankfurt.de/biodiv?lang=en>
- further digitalisation activities
(e.g. “Sammlung Deutscher Drucke 1801-1870”)

... whenever possible, content shall also be incorporated in BHL

Building a BIOfid corpus

One of BIOfid's aims: extraction of biodiversity data from literature (focus Central Europe, vascular plants, birds, butterflies)

-> enabling more efficient IT-based analyses

Corpus will be assembled from various sources:

- digitised content from UB JCS
- content digitised elsewhere but *lacking* free licenses
- content digitised elsewhere *with* free licenses / PD
- born-digital content from OA journals

Domain-specific ontologies

Ontologies are needed as a basis for semantically enhancing the literature, but

- not enough depth in existing ontologies
- not enough coverage of German terms in existing ontologies

-> development and expansion of ontologies as a BIOfid task

some examples:

- [LepAO - Lepidoptera Anatomy Ontology](#)
- [ChilAO - Chilopoda Anatomy Ontology](#)
- [GTO - GBIF Taxonomy Ontology](#)
- [ENVO - Environment Ontology](#)

Further BIOfid services

There's more... (but not part of today's workshop):

- open access journal platform (OJS)
- digitisation on demand (including copyright clearance for orphaned works)
- supply of printed literature via document delivery or ILL
- FID licenses for electronic content (restricted to Germany: *JSTOR Global Plants; Birds of the World; Brill e-books*)

Workshop Agenda

1. Introduction to BIOfid
- 2. BIOfid Web-Interface**
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

BIOfid Search Web-Interface

<https://www.biofid.de/search/>




Beta Version!

The search includes the organismic groups of birds, butterflies, and vascular plants.

Currently, fulltexts from the Digital Collection Biology and a selection of fulltexts from Zobodat are indexed. Fulltexts from BHL are in preparation.

Keyword
 Semantic

supported by data from:

Example queries

- Suche nach Orten:
 - [Fagus in Hessen](#)
 - [Sauergräser in Schwaben](#)
- Differenzierung zwischen "und" und "oder" Anfragen:
 - [Schmetterlinge und Vögel](#)
 - [Schmetterlinge oder Vögel](#)
 - [Fagus sylvatica, Weißstorch und Eichen](#)




Restricted taxa only!

Example Queries
(only in German)

BIOfid Search Web-Interface

Taxus baccata in Berlin

Keyword Semantic

supported by data from:
  

Recognized search terms
 Taxus baccata in Berlin
 PLANT LOCATION

Found Taxa

Search Results
 Max. 1000 Data

Filter Search

Filter
 Database (Un-)select all
 Dig. Samm. UB JCS (30)
 Annotated Full Text
 Available (30)

Document
 30 Search Results
 Number of displayed hits: 10

Botanische Exkursionen und pflanzengeographische Studien in der Schweiz: VI. Allgemeines
 Download as PDF Citation
 Database: Dig. Samm. UB JCS Language: Unbekannt Publication year: 1905 Author: Unknown Journal: Botanische Exkursionen und pflanzengeographische Studien in der Schweiz

Botanische Exkursionen und pflanzengeographische Studien in der Schweiz: V. Die einzelnen Fundorte mit ihren Resten
 Download as PDF Citation
 Database: Dig. Samm. UB JCS Language: Unbekannt Publication year: 1905 Author: Unknown Journal: Botanische Exkursionen und pflanzengeographische Studien in der Schweiz

Natural Language Search

What was actually recognized

Get the text annotations

BIOfid Search Web-Interface

Search

Keyword Semantic

supported by data from:

Recognized search terms

Taxus baccata

in

Berlin

PLANT
LOCATION

Search Results

Max. 1000 Data

Download ▾

Document

30 Search Results Number of displayed hits: 10 ▾

Botanische Exkursionen und pflanzengeographische Studien in der Schweiz: VI. Allgemeines

Download as PDF Citation

Database: Dig. Samm. UB JCS Language: Unbekannt Publication year: 1905 Author: Unknown Journal: Botanische Exkursionen und pflanzengeographische Studien in der Schweiz

▼ Page Hits

Seite 96 (2 Hits):

...96 E. Neuweiler. | I IN **Kreuzberg**: (**Betula**) (**Prunus**) (**Convolvulus**) (**Prunus avium**) (**Prunus insititia**) (**Prunus domestica**) ...

... (**Panicum miliaceum**) oder (**Setaria italica**) (**Polygonum**) (**Convolvulus**) Polchblech: (**Taxus baccata**) (**Mainz**): Anomodon curlipendulum, Braohythecium rulabulum, Camptothecium lutescens, (**Hypnum lutescens**) (**Hypnum**) pseudotamariscinum, (**Hypnum**) Schre-beri, (**Hypnum splendens**) (**Mnium**) punctatum. ...

Seite 99 (2 Hits):

...7. – Über die im **königlichen Museum** zu Berlin aufbewahrten Pflanzenreste aus altägyptischen Gräbern. Ethnogr. ...

...**Buchholz**: Fund eines Leinsamenvorrats in den Oberresten einer prähistorischen Wohnstätte bei **Frehne**, Kreis Ostpriegnitz. Ethnogr. ...

Found Taxa

Filter Search

Reset all filters

Filter

Database (Un-)select all

Dig. Samm. UB JCS (30)

Annotated Full Text

Available (30)

Publication year (Un-)select all

1850 - 1859 (1)

1870 - 1879 (3)

Link to the digitized page




Interactive annotations!!

BIOfid Search Web-Interface

Search

Keyword Semantic

supported by data from:

Recognized search terms

Taxus baccata

in

Berlin

PLANT
LOCATION

Search Results

Max. 1000 Data

Download ▾

Number of displayed hits:

Filter Search

Reset all filters

Filter

Database (Un-)select all

Dig. Samm. UB JCS (30)

Annotated Full Text


Available (30)

Publication year (Un-)select all

1850 - 1859 (1)

1870 - 1879 (3)

Bei Wikipedia nachschlagen



Kreuzberg

Kreuzberg is a district of Berlin, Germany. It is part of the Friedrichshain-Kreuzberg borough located south of Mitte. During the Cold War era, it was one of the poorest areas of West Berlin, but since German reunification in 1990 it has become more gentrified and known for its arts scene.

Author: Unknown Journal: Botanische Exkursionen und pflanzengeographische

...96 E. Neuweiler. | IIN **Kreuzberg**. (**Polygonum**) Gonvolvulus, Prunus avium, (**Prunus insititia**), (**Prunus domestica**). ...

...(**Panicum miliaceum**) oder (**Setaria italica**), (**Polygonum**) Convolvulus. Polchblech: (**Taxus baccata**). (**Mainz**): Anomodon curlipendulum, Braohythecium rulabulum, Camptothe- cium lutescens, (**Hypnum lutescens**), (**Hypnum**) pseudotamariscinum, (**Hypnum**) Schre- beri, (**Hypnum splendens**), (**Mnium**) punctatum. ...

Seite 99 (2 Hits):

...7. – Über die im **königlichen Museum** zu Berlin aufbewahrten Pflanzenreste aus altägyptischen Gräbern. Ethnogr. ...

...**Buchholz**: Fund eines Leinsamenvorrats in den Oberresten einer prähistorischen Wohnstätte bei **Frehne**, Kreis Ostpriegnitz. Ethnogr. ...

Found Taxa

09.06.2022 Workshop at SPNHC 2022

BIOfid Search Web-Interface

Taxus baccata in Berlin

Keyword Semantic

supported by data from:

Recognized search terms

Taxus baccata in Berlin
PLANT LOCATION

Search Results
 Max. 1000 Data

Filter Search

Database (Un-)select all
 Dig. Samm. UB JCS (30)

Annotated Full Text
 Available (30)

Publication year (Un-)select all
 1850 - 1859 (1)
 1870 - 1879 (3)

Document

30 Search Results

Botanische Exkursionen und pflanzengeographische

Download as PDF Citation

Database: Dig. Samm. UB JCS Language: Unbekannt Publication year: ...

Studien in der Schweiz

▼ Page Hits

Seite 96 (2 Hits):

...96 E. Neuweiler. | I IN **Kreuzberg** (**Polygonum**) Gonvo...

...(**Panicum miliaceum**) oder (**Setaria italica**) (**Polygonum**) Convolvulus. Polchblech: (**Taxus baccata**). (**Mainz**): Anomodon curlipendulum, Braohythecium rulabulum, Camptothecium lutescens, (**Hypnum lutescens**), (**Hypnum**) pseudotamariscinum, (**Hypnum**) Schre-beri, (**Hypnum splendens**) (**Mnium**) punctatum. ...

Seite 99 (2 Hits):

...7. — Über die im **königlichen Museum** zu Berlin aufbewahrten Pflanzenreste aus altägyptischen Gräbern. Ethnogr. ...

...**Buchholz**: Fund eines Leinsamenvorrats in den Oberresten einer prähistorischen Wohnstätte bei **Frehne**, Kreis Ostpriegnitz. Ethnogr. ...

...

Bei Wikipedia nachschlagen
 BIOfid Taxon Daten
 GBIF Species Seite
 Suche in BIOfid-Portal

graphische

Found Taxa

Annotation context information

BIOfid URI

BIOfid Search Web-Interface

BIOfid URI: <https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517>

Group	Property	Value
Taxus baccata L.		
Systematics	Kingdom	Plantae
	Phylum	Tracheophyta
	Class	Pinopsida
	Order	Pinales
	Family	Taxaceae
	Genus	Taxus
	Scientific Name	Taxus baccata L.
	Name Author	L.
	Label	Taxus baccata
	Taxon Rank	species (en)
	Vernacular Name	Eibe (de)
	Vernacular Name	Beereneibe (de)
	Vernacular Name	Europäische Eibe (de)
	Vernacular Name	Gemeine Eibe (de)
	Vernacular Name	Gewöhnliche Eibe (de)
	Vernacular Name	yew (en)
	Vernacular Name	Common yew (en)
	Vernacular Name	English yew (en)
	Vernacular Name	European yew (en)
	Vernacular Name	Yew (en)
	Vernacular Name	if (fr)
	Vernacular Name	if commun (fr)
Taxonomic Status	accepted	
Parent Name	Taxus	
Traits	flower color	yellow
Metadata	TRY-Species ID	53526
Misc	Type	GBIF: Taxon
	Is in BIOfid Literature Corpus	True
	Type	flower yellow
	Type	Named Individual
	Type	Named Individual

What your browser sees

```

title: "Taxus baccata L."
uri: "https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517"
data:
  0:
    predicate:
      type: "uri"
      value: "https://www.biofid.de/bio-ontologies#has_petal_color"
    object:
      type: "uri"
      value: "http://purl.obolibrary.org/obo/PATO_0000324"
    objectLabel:
      type: "literal"
      value: "yellow"
    predicateLabel:
      type: "literal"
      value: "flower color"
  1:
    predicate:
      type: "uri"
      value: "https://dwc.tdvw.org/terms/#class"
    object:
      type: "uri"
      value: "https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/194"
    objectLabel:
      type: "typed-literal"
      datatype: "http://www.w3.org/2001/XMLSchema#string"
      value: "Pinopsida"
  
```

What your crawler sees






BIOfid Search Web-Interface

Taxus baccata in Berlin

Suchen

Stichwort Semantisch

unterstützt durch Daten von:

Erkannte Suchbegriffe

Taxus baccata

in

Berlin

PFLANZE
ORT

Suchergebnisse

Max. 1000 Datensätze

Download ▾

CSV

JSON

Alle Filter zurücksetzen

Filtern

Datenbank Alle an/aus

Dig. Samm. UB JCS (30)

Annotierter Volltext

Verfügbar (30)

Publikationsjahr Alle an/aus

1850 - 1859 (1)

1870 - 1879 (3)

Dokumente

30 Suchergebnisse
Angezeigte Trefferzahl: 10 ▾

Botanische Exkursionen und pflanzengeographische Studien in der Schweiz: VI. Allgemeines

Download als PDF **Zitation**

Datenbank: Dig. Samm. UB JCS Sprache: Unbekannt Publikationsjahr: 1905 Autor: Unbekannt Zeitschrift: Botanische Exkursionen und pflanzengeographische Studien in der Schweiz

▼ Seitentreffer

Seite 96 (2 Treffer):

...96 E. Neuweiler. | IN **Kreuzberg**: **Polygonum** Gonvolvulus, Prunus avium, **Prunus insititia**, **Prunus domestica**, ...

...**Panicum miliaceum** oder **Setaria italica**, **Polygonum** Convolvulus. Polchblech: **Taxus baccata**. **Mainz**: Anomodon curlipendulum, Braohytheicum rulabulum, Camptothecium lutescens, **Hypnum lutescens**, **Hypnum** pseudotamariscinum, **Hypnum** Schre-beri, **Hypnum splendens**, **Mnium** punctatum. ...

Seite 99 (2 Treffer):


...7. – Über die im **königlichen Museum** zu Berlin aufbewahrten Pflanzenreste aus altägyptischen Gräbern. Ethnogr. ...

...**Buchholz**: Fund eines Leinsamenvorrats in den Oberresten einer prähistorischen Wohnstätte bei **Frehne**, Kreis Ostpriegnitz. Ethnogr.

...

Gefundene Taxa

Download annotations



BIOfid JSON Download Data

```

...{
  "Authors": "Unbekannt",
  "Journal": "Botanische Exkursionen und pflanzengeographische Studien in der Schweiz",
  "PublicationYear": 1905,
  "Source": "",
  "TextExtracts": [
    {
      "Label": "Seite 96",
      "Page URL": "https://sammlungen.ub.uni-frankfurt.de/4439682?query=taxus%20baccata%20kreuzberg",
      "TextAnnotations": [
        "<em class='location_place' id='15c6369f95' wikidata='http://www.wikidata.org/entity/Q308928'>Kreuzberg</em>",
        "<em biofid-uri-0='https://www.biofid.de/bio-ontologies/Tracheophyta#GBIF_5284517'"
        biofid-uri-1='https://www.biofid.de/bio-ontologies/Tracheophyta#GBIF_8235017'"
        biofid-uri-2='https://www.biofid.de/bio-ontologies/Tracheophyta#GBIF_8343790' class='taxon facet biofid-included' id='6b204fa5c2'"
        wikidata='http://www.wikidata.org/entity/Q179729'>Taxus baccata</em>"
      ]
    },
    {
      "Label": "Seite 99",
      "Page URL": "https://sammlungen.ub.uni-frankfurt.de/4439685?query=buchholz%20k%C3%B6niglichen%20museum",
      "TextAnnotations": [
        "<em class='location_place' id='4b084ae4ac' wikidata='http://www.wikidata.org/entity/Q156722'>k\u00f6niglichen Museum</em>",
        "<em class='location_place' id='22d5d33242' wikidata='http://www.wikidata.org/entity/Q675015'>Buchholz</em>"
      ]
    }
  ],
  "Title": "VI. Allgemeines",
  "URL": "http://sammlungen.ub.uni-frankfurt.de/botanik/periodical/pageview/5027912"
}

```

BIOfid Text Annotation Demo Part I

Google Collab: <https://t1p.de/biofid-annotation>

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
- 3. Natural Language Processing-Pipeline (Hands-on)**
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

Demo

1. Go to <http://spnhc2022.texttechnologylab.org> and upload text file



The screenshot shows the interface of the SPNHC 22 Workshop. At the top left, it displays 'Fachinformationsdienst Biodiversitätsforschung' with a green leaf logo. To the right, it says 'SPNHC 22 Workshop' and features a blue portrait of a woman next to the text 'text technology'. Below this, there is a white bar with three buttons: 'Datei auswählen', 'Keine ausgewählt', and 'Process Text'. The background is a solid green color.

Demo

Example file:

<https://sammlungen.ub.uni-frankfurt.de/botanik/content/rpage/4162821>

I am pleased to think that the study of the Rhizocephala began in Naples in 1787, when Cavolini (1) published his volume "Sulla generazione dei Pesci e dei Granchi". Although his view of the adult Sacculina as a morbid growth produced by the crab to surround a deposit of the eggs of a parasite is only of historic interest, we may admire the fortune and merit of the Neapolitan naturalist in recognizing the Crustaceous nature of the parasite through its Nauplius larvae, and also in discovering the Cryptoniscus larva of a parasite of the Sacculina, which he rightly referred to the Isopoda, under the name *Oniscus squilliformis*. His work was ignored till 1854, and unsurpassed till 1862.

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. **XML**
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

XML

1. Understanding the output file: Stand-off annotation
2. Separating primary data (“the text”) from annotations. Why?
3. Inline annotation and discontinuous constituents
 E.g. “look up” is an English verb, but: “I looked it up.”

```

      <pos=noun>I</pos>
      <pos=verb>looked</pos>
      <pos=pron>it</pos>
      <pos=??>up</pos> (?? could be verb prefix, but connection to verb
      lost [e.g., if more than one verb in sentence])
      <pos=punct>.</pos>
      
```
4. Related problem: inline annotation and several annotations for the same input → “crossing annotation spans”

XML

1. Understanding the output file: Markables, Indices, Features

markable:	I		a	m		p	l	e	a	s	e	d		t	o		
index:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	s0		s2		start=5												
	e1		e4		end=12												
feature:	<pre><pos value="ppron" begin="0" end="1" id="100"/> <lemma value="be" begin="2" end="4" id="201"/> <verb_complex value="be pleased" begin="2" end="12" id="990"/></pre>																

XML

1. Different ways and formats, we use UIMA CAS and XMI
(<https://uima.apache.org/d/uimaj-current/apidocs/org/apache/uima/cas/CAS.html>)
2. *Unstructured Information Management Architecture (UIMA)* , includes the
 - CAS (*Common Analysis System*) for creating and handling the data that annotators manipulate
 - CAS includes the type system which defines two kinds of entities, *types* and *features*.
3. Open standard format: XMI (*XML Metadata Interchange*)
4. Let's have a look at the sections of the output file

XML

1. Loading name spaces to regiment elements in a unique way
`<xmi:XMI xmlns:xmi="http://www.omg.org/XMI" ...`
2. The “document element” (near top): `<tcas:DocumentAnnotation begin="0" end="691" language="en" sofa="1" xmi:id="8"/>`
3. ...and its link to the input string (close to bottom): `<cas:Sofa mimeType="text" sofaID="_InitialView" sofaNum="1" sofaString="I am pleased to think ...`
4. ... which is our *Subject of Analysis* (**sofa**)

XML

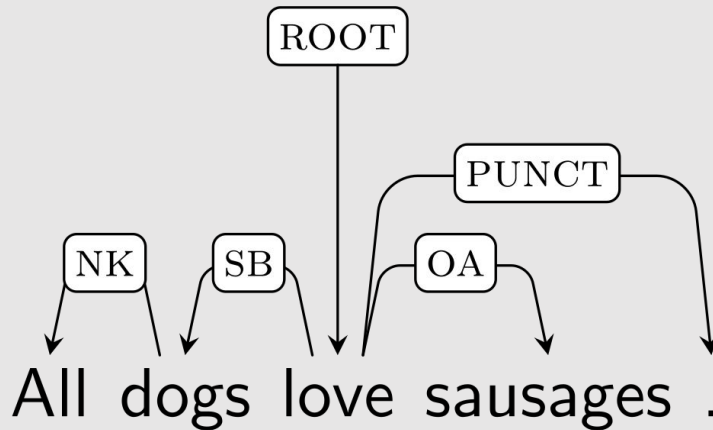
1. Annotations are added in terms of the *Markable–Index–Feature-Model*
2. Tokens (i.e., “single words”): `<type6:Token xmi:id="165" sofa="1" begin="16" end="21" lemma="4423" pos="2737" morph="6048" order="0"/>`
 - This is “think”, begins at index 16, ends at index 21 and belong to sofa 1 (our document)
 - important: “think” becomes unique with the ID “189”
 - it is also linked to other feature layers: lemma, morph(ology), part of speech (pos)
3. So, what’s “think”’s part of speech? Follow the pos-ID: `<pos:POS_VERB xmi:id="2737" sofa="1" begin="16" end="21" PosValue="VB" coarseValue="VERB"/>`

XML

1. Various layers build according to this model: besides lemmas, morphology and POS, the single sentences are distinguished and receive a syntactic annotation in terms of dependency structure.
2. E.g.: `<type6:Sentence begin="632" end="684" sofa="1" xmi:id="73"/>`
3. “I am pleased to think [...]”: “think” depends on “pleased”, which is the governor in syntax (ID 149): `<dependency:Dependency xmi:id="9685" sofa="1" begin="16" end="21" Governor="125" Dependent="165" DependencyType="XCOMP" flavor="basic"/>`

Excursus: Dependency Structure

1. Dependency: syntactic relation between *words*: a **governor** or head and its **dependent**
2. Binary tree with labelled edges
3. Simple example: *All dogs love sausages.*



- Root: head of sentence (main verb)
- SB: subject
- OA: accusative object
- NK: noun kernel (part of noun phrase)
- PUNCT: punctuation

Named Entities

1. Finally: Taxa!
2. *Global Names Finder* (gnfinder; Mozzherin, Myltsev & Zalavadiya, <https://zenodo.org/record/6537485>) und "*Taxon Gazetteer*" (Ahmed, Stoeckel, Driller, Pachzelt & Mehler, DOI: 10.18653/v1/K19-1081)
3. "Sacculina":
 - Gazetteer: `<type19:Taxon xmi:id="12004" sofa="1" begin="197" end="206" value="http://www.wikidata.org/entity/Q150610, https://www.biofid.de/bio-ontologies/Tracheophyta%23GBIF_7297853"/>`
 - gnfinder: `<type19:Taxon xmi:id="12318" sofa="1" begin="197" end="206" value="Sacculina" identifier="https://www.catalogueoflife.org/data/taxon/7BHH"/>`

Named Entities

1. Finally: T
2. *Global N*
<https://z>
"Taxon G
10.18653
3. "Sacculin
 - *Gazet*
begin
value
[https](https://z)
72978
 - *gnfin*
begin
ident
HH"/>



- [Main page](#)
- [Community portal](#)
- [Project chat](#)
- [Create a new Item](#)
- [Recent changes](#)
- [Random item](#)
- [Query Service](#)
- [Nearby](#)
- [Help](#)
- [Donate](#)

- [Lexicographical data](#)
- [Create a new Lexeme](#)
- [Recent changes](#)
- [Random Lexeme](#)

- [Tools](#)
- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Permanent link](#)
- [Page information](#)
- [Concept URI](#)
- [Cite this page](#)

[English](#) [Not logged in](#) [Talk](#)

Item
Discussion
Read
[View history](#)

Sacculina (Q150610)

genus of crustaceans [edit](#)

[Wikipedia](#) (17 entries) [edit](#)

Language	Label	Description	Also known as
English	Sacculina	genus of crustaceans	
German	Sacculina	Gattung der Familie Sacculinidae (Sacculinidae)	
French	Sacculina	genre de crustacés	Sacculine
Bavarian	No label defined	No description defined	

[In more languages](#)

Statements

instance of taxon 0 references

image depicts

[Sacculina carcini.jpg](#)
1,874 × 1,235; 728 KB

[depicts](#) [Sacculina carcini](#)

Catalogue of Life

Sacculina

COL Identifier	7BHH ⓘ
Name	<i>Sacculina</i>
Checklist status	accepted genus
Nomenclatural Status	potentially valid
Classification	unranked: <i>Biota</i> > kingdom: <i>Animalia</i> > phylum: <i>Arthropoda</i> > class: <i>M</i>
Taxonomic scrutiny	CoL

Views

1. *Views*: list of elements that belong to a “view”: `<cas:View members="8
13 25 37 49 61 ...`
2. This way, different “views” (e.g. annotations) can be associated with the same source text
3. Single annotators can also be tracked via *fingerprints* (home-made function)

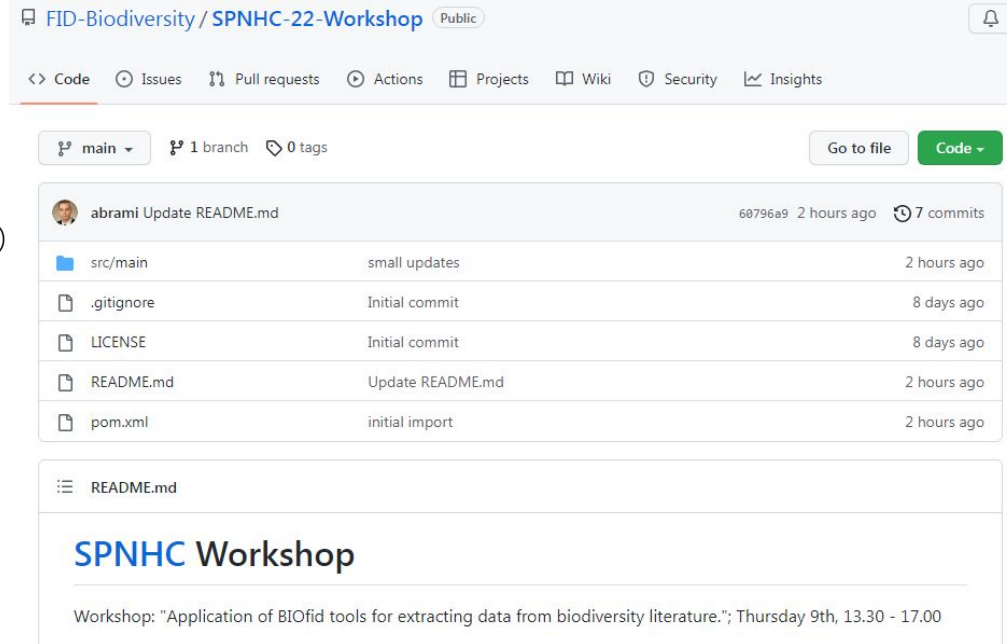
Conclusion: Versatile, flexible and portable format

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
- 5. Java-Pipeline**
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

Java pipeline

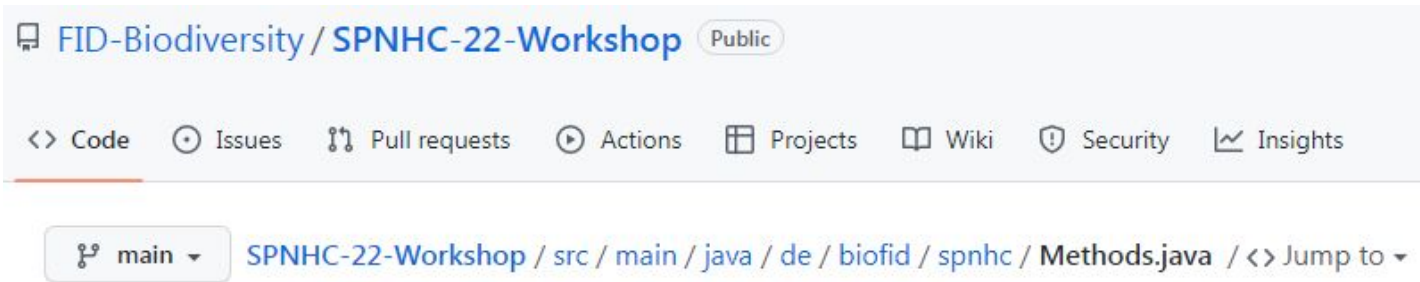
1. Website-Pipeline is available as github project at <https://github.com/FID-Biodiversity/SPNHC-22-Workshop>
2. requires Java SDK version 11
3. clone from github
4. open in IntelliJ (<https://www.jetbrains.com/idea/>)
5. build your own pipeline



The screenshot shows the GitHub repository page for 'FID-Biodiversity / SPNHC-22-Workshop'. The repository is public and has 1 branch (main) and 0 tags. The commit history shows a recent update to README.md by 'abrami' 2 hours ago, with 7 commits. The file list includes src/main, .gitignore, LICENSE, README.md, and pom.xml. The README.md file is open, showing the title 'SPNHC Workshop' and the description: 'Workshop: "Application of BIOfid tools for extracting data from biodiversity literature."; Thursday 9th, 13.30 - 17.00'.

Java pipeline

1. The workshop project provides four tools: SpaCy, BIOfid's HeidelTime extension, GeoNames tagger, GNFinder
2. Go to Methods.java:



3. Select the tools you want (line 76):

```
75
76     pCas = processNLP.process(sContent, NLP.getSpacy(), NLP.getHeidelTimeBioFID(), NLP.getGeoNames(), NLP.getGNFinder());
77
```

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
- 6. – Break –**
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
10. Discussion

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
- 7. Colab (Hands-on)**
8. Semantic Role Labeling
9. Survey
10. Discussion

BIOfid Text Annotation Demo Part II

Google Collab: <https://t1p.de/biofid-annotation>

BIOfid API - Context Analysis

<https://www.biofid.de/api/v1/getTermContext?term=https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517>

Returns the most common co-occurrences for taxa and locations in the text:



BIOfid API - Context Analysis

<https://www.biofid.de/api/v1/getTermContext?term=https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517>

Returns the most common co-occurrences for taxa and locations in the text:

Taxa

1.	<i>Taxus baccata</i> L.	(278 cooccurrences)
2.	<i>Taxus baccata</i> Thunb.	(270)
3.	<i>Taxus baccata</i> Hook.	(270)
4.	Plantae	(220)
...		
7.	<i>Quercus</i> L.	(122)
8.	<i>Carex</i> L.	(118)
...		
32.	<i>Daphne mezereum</i> L.	(99)
...		
51.	<i>Juniperus communis</i> L.	(87)
...		
73.	<i>Viburnum opulus</i> L.	(84)



BIOfid API - Context Analysis

<https://www.biofid.de/api/v1/getTermContext?term=https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517>

Returns the most common co-occurrences for taxa and locations in the text:

Taxa

1.	<i>Taxus baccata</i> L.	(278 cooccurrences)
2.	<i>Taxus baccata</i> Thunb.	(270)
3.	<i>Taxus baccata</i> Hook.	(270)
4.	Plantae	(220)
...		
7.	<i>Quercus</i> L.	(122)
8.	<i>Carex</i> L.	(118)
...		
32.	<i>Daphne mezereum</i> L.	(99)
...		
51.	<i>Juniperus communis</i> L.	(87)
...		
73.	<i>Viburnum opulus</i> L.	(84)

Locations

1.	Central Eastern Alps	(6)
2.	Limestone Alps	(5)
3.	Ford	(5)
4.	Corridor	(5)
5.	Left Bank of the Rhine	(5)
6.	Lesnoje (Russia)	(5)
7.	Zirl (Austria)	(5)
8.	South West Germany	(5)
9.	Saint Pierre and Miquelon	(5)
10.	Komsomolsk (Russia)	(5)
11.	...	



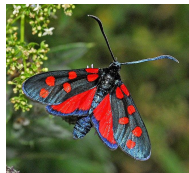
GBIF Data 1819-1920

BIOfid API - Context Analysis

Base URL: <https://www.biofid.de/api/v1/getTermContext>

Ways to query

By String Term



<https://www.biofid.de/api/v1/getTermContext?term=zygaenidae>

By Wikidata URI



<https://www.biofid.de/api/v1/getTermContext?term=http://www.wikidata.org/entity/Q980>

By BIOfid URI

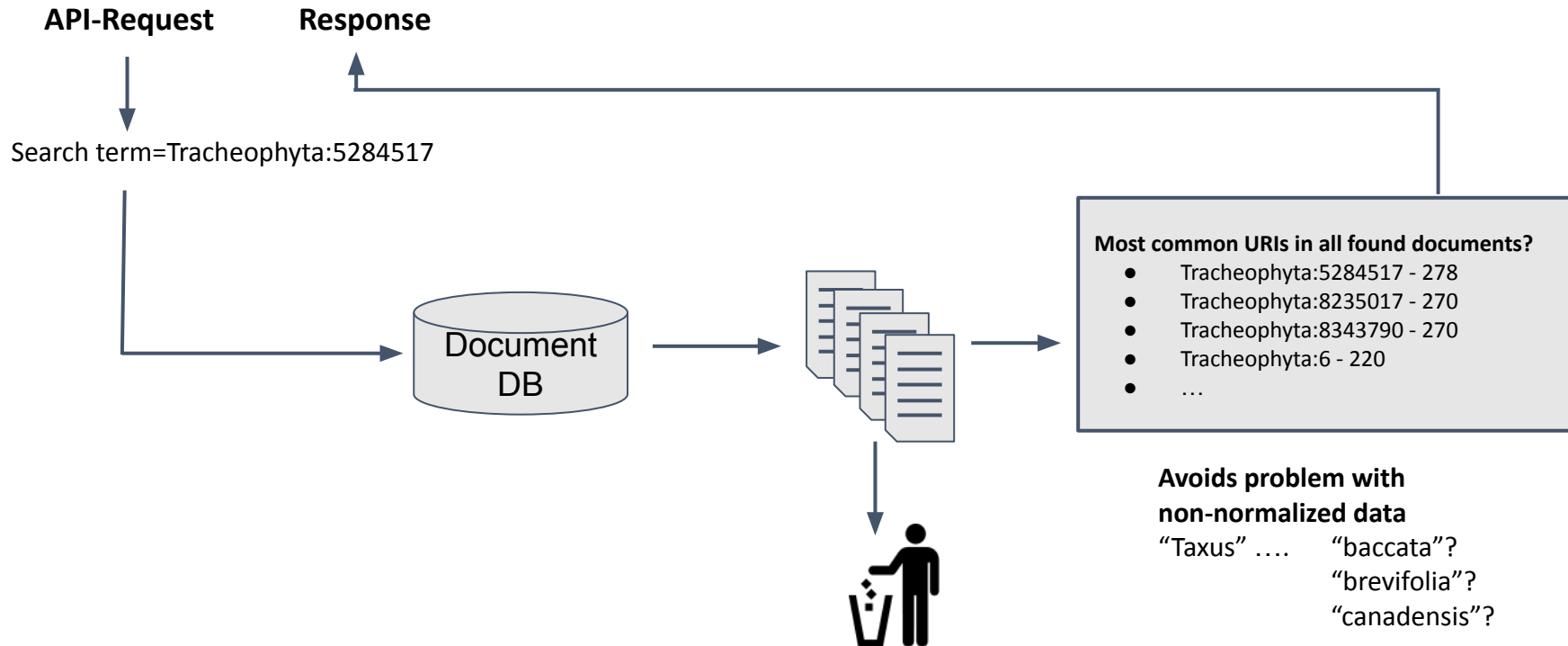


<https://www.biofid.de/api/v1/getTermContext?term=https://www.biofid.de/bio-ontologies/Tracheophyta/gbif/5284517>

[Taxus baccata](#) by [Didier Descouens](#) under [CC BY-SA 4.0](#)

[Zygaena transalpina](#) Val Veny, Aosta, Italy by [Hectonichus](#) under [CC BY-SA 4.0](#)

BIOfid API - Context Analysis



BIOfid API - Context Analysis

Advantages of the Context Analysis

- Aggregate the information pieces (“signals”) from hundreds of manuscripts
- Although a specific interaction is never mentioned, the CA reveals it

The power of normalized data!

Make a Context Analysis yourself

Google Collab: <https://t1p.de/biofid-context>

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
- 8. Semantic Role Labeling**
9. Survey
10. Discussion

SRL: Semantic Role Labeling

1. From named entities to **relations**, and in a systematic manner.
2. Approach: assigning participant roles to syntactic arguments.
3. Simple example (repeated): *All dogs love sausages.*
 - **Relation**: *loving*
 - **“Lover”**: *all dogs*
 - **“Loved”**: *sausages*
4. **Not**: *Sausages love all dogs.*

SRL

1. Representations in computational linguistics:

- Positional: $\text{love}(\text{all dogs}, \text{sausages}) \neq \text{love}(\text{sausages}, \text{all dogs})$
- Verb specific: $[\text{relation} = \text{love}, \text{lover} = \text{all dogs}, \text{loved} = \text{sausages}]$
- General: $[\text{relation} = \text{love}, \text{agent} = \text{all dogs}, \text{patient} = \text{sausages}]$
- Abstract: $[\text{relation} = \text{love}, \text{ARG0} = \text{all dogs}, \text{ARG1} = \text{sausages}]$

2. PropBank (Bonial et al. 2015): abstract approach

ARG0	proto-agent	ARG3	starting point, benefactive, attribute
ARG1	proto-patient	ARG4	ending point
ARG2	instrument, benefactive, attribute	modifiers	modals, location, time, negation, purpose, ...

SRL: (Dependency) Syntax is not enough

1. Real world example:

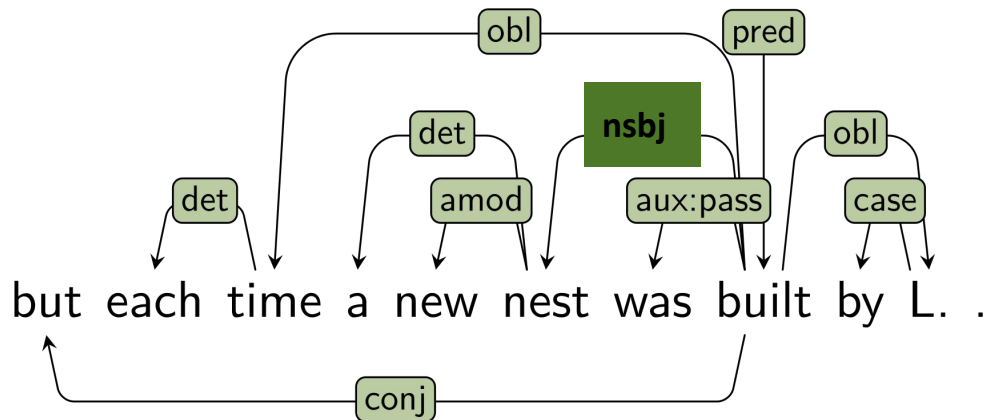
“[...] jedesmal aber wurde ein neues Nest gebaut” [von *Lantus exubitor L.*] (BIOFid 4523122)

[...] *but each time a new nest was built [by Lantus exubitor L.]*

2. Right: Dependencies (UD)

3. Roles:

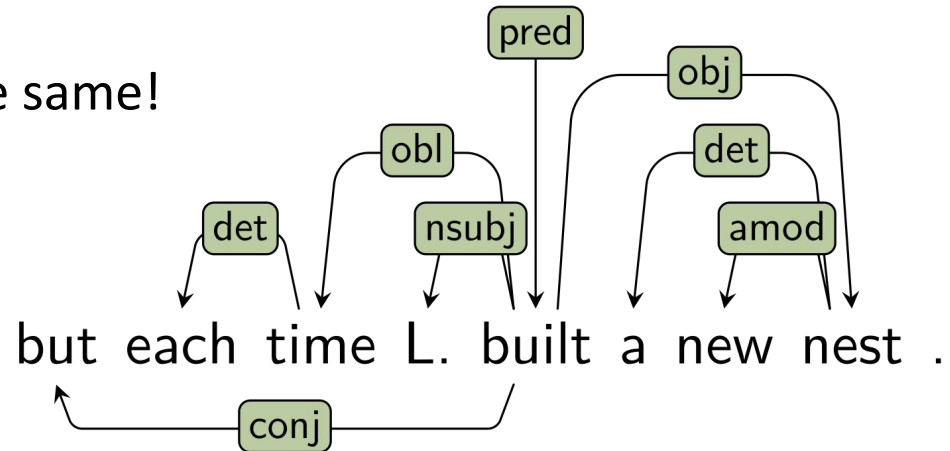
- **relation**: *building*
- **builder**: *Lantus exubitor L.* (oblique argument)
- **built**: *new nest* (direct object)



<https://universaldependencies.org>, UD

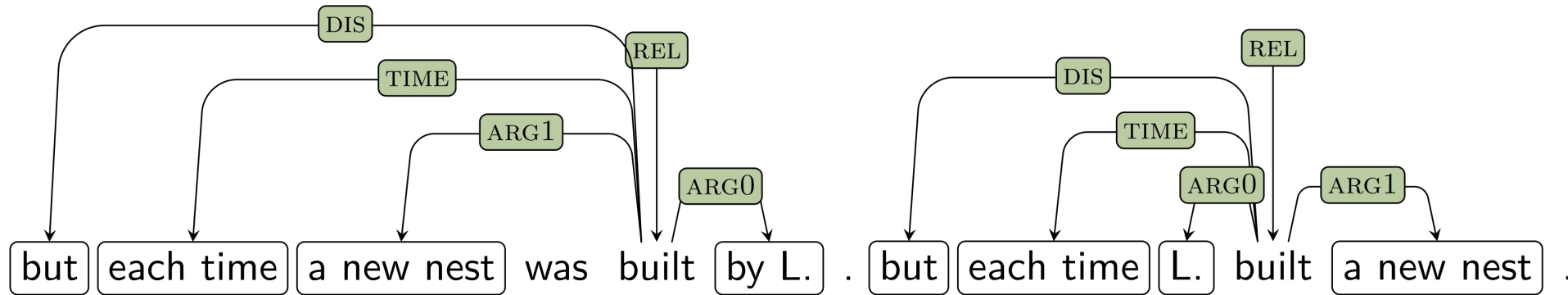
SRL: (Dependency) Syntax is not enough

1. Active voice: *jedesmal aber baute L. ein neues Nest.*
but each time L. built a new nest
2. quite different syntactic structure: *Lantus exubitor L. is*
n(ominal)subj(ect)
3. .. but semantic roles remain the same!



SRL

1. example annotations for our sample sentence
2. But: works reasonably well for English, but not (yet) for German -> **ongoing work for next edition**



Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
- 9. Survey**
10. Discussion

BIOfid-Survey

<https://t1p.de/biofid-survey>

Workshop Agenda

1. Introduction to BIOfid
2. BIOfid Web-Interface
3. Natural Language Processing-Pipeline (Hands-on)
4. XML
5. Java-Pipeline
6. – Break –
7. Colab (Hands-on)
8. Semantic Role Labeling
9. Survey
- 10. Discussion**