Passage retrieval services: A RESTful API, which will input a set of keywords and return a set of articles/ passages ranked by relevance

Deliverable D11.3

31 October 2022

Authors

Emilie Pasche, Julien Gobeill, Alexandre Flament, Jeevanthi Liyanapathirana,

Pierre-André Michel, Déborah Caucheteur, Esteban Gaillac, Nona Naderi, Patrick Ruch

SIB Swiss Institute of Bioinformatics & HES-SO

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Passage retrieval services: A RESTful API, which will input a set of keywords and return a set of articles/ passages ranked by relevance
Deliverable n°:	D11.3
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP11
Lead beneficiary:	SIB
Citation:	Pasche, E., Gobeill, J., Flament, A., Liyanapathirana, J., Michel, P., Caucheteur, D., Gaillac, E., Naderi, N. & Ruch, P.. (2022). *Passage retrieval services: A RESTful API, which will input a set of keywords and return a set of articles/ passages ranked by relevance*. Deliverable D11.3 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 18
Actual submission date:	31 October 2022

Deliverable status:

Version	Status	Date	Author(s)
1.0	Reviewed by Sharif Islam Mathias Dillen, Jose Benito Gonzales Lopez.	27 Oct 2022	WP11 members
0	Initial report	12 Oct 2022	See authors list.

The content of this deliverable does not necessarily reflect the official opinions of the European Commission or other institutions of the European Union.

Table of contents

Executive summary

Background: The BiCIKL project is born from the vision that biodiversity data are most useful if they are presented as a nexus of data that can be integrated and viewed from different starting points. BiCIKL's goal is to realise that vision by linking biodiversity data, in particular literature, molecular sequences, specimens, nomenclature and analytics. To do so, we need to better understand the existing infrastructures, their limitations, the nature of the data they hold, the services they provide and particularly how they can interoperate.

Objectives: WP11 aims at exploring solutions to deliver a FAIR Data Place, i.e. a one stop service to help members of the biodiversity community to navigate the constellation of biodiversity databases. This deliverable aims at describing the design of WP11's passage retrieval services. The deliverable is a demonstrator and this report provides a synthetic description of the developments. While the main achievement is the API, the report also drafts how the services will be integrated in subsequent WP11 deliverables (D11.4-5, i.e. the Search and Question-Answering Portal).

Methods: The main development steps are the following: 1) acquisition of biodiversity contents from Plazi for harvesting, semantic enrichment and indexing in the SIB Literature Services to create the largest machine readable biodiversity library; 2) development of biodiversity-specific annotation workflows to deliver powerful text analytics in the domain of biotic interactions, as defined in D11.1; 3) development of an original graphical user interface to explore the biotic universe; 4) evaluation of the passage retrieval services applied to a subset of biotic interactions as defined in D11.1.

Results: We report on results according to different dimensions: 1) Architecture of services, 2) Application Programming Interface, 3) User interface, 4) Passage retrieval efficiency and finally, 5) Passage retrieval effectiveness. For the latter results, it is reported that normalization of species and ROBI interactions result in an improved retrieval effectiveness.

Conclusion: The passage retrieval service is ready, together with the prototype of the Graphical User Interface. The service will serve as the basis for D11.4, the question answering services. Further works are needed to improve the response time of the service, up to the point where user interactions are fluent.

1. Introduction

This deliverable capitalises on use cases described early in D11.1 Evaluation benchmark: a database of questions associated with a set of relevant articles (e.g. PMIDs) to measure the progress of the WP towards an effective search service , where a set of questions has been collected. The questions were provided with a set of possible answers together with an evidence link (e.g. publication or database) likely to justify the relevance of the answer. Because more than half of the questions are directly related to biotic interactions, it was concluded in this first deliverable that WP11 will tentatively focus on these types of questions, see example in Appendix A of D11.1.

The deliverable is also directly dependent on T11.2 and D11.2 Search and link association services: A RESTful API, which will input a link/ accession number and return a ranked list of neighbors links with a confidence score, which aims at designing machine learning services that suggest new associations between biodiversity entities, and in particular biotic interactions between species. Both D11.2 and D11.3 will serve a shared resource, the so-called FAIR Data place. D11.2 is also developing a graph visualisation interface, which could also be used to visualise entities and relationships identified from the literature, as returned by D11.3 and soon D11.4. Last, but not least, the deliverable is also connected with WP6, which aims at structuring/enriching literature contents, and the BKH (Biodiversity Knowledge Hub), as we are monitoring the progress of BICIKL's RDF repositories, such as OpenBioDiv, which could potentially be leveraged to answer some of the D11.1 questions using the SPARQL queries.

2. Methodology

In this section, we detail the following items:

1) acquisition of biodiversity contents from Plazi for harvesting, semantic enrichment and indexing in the SIB Literature Services,

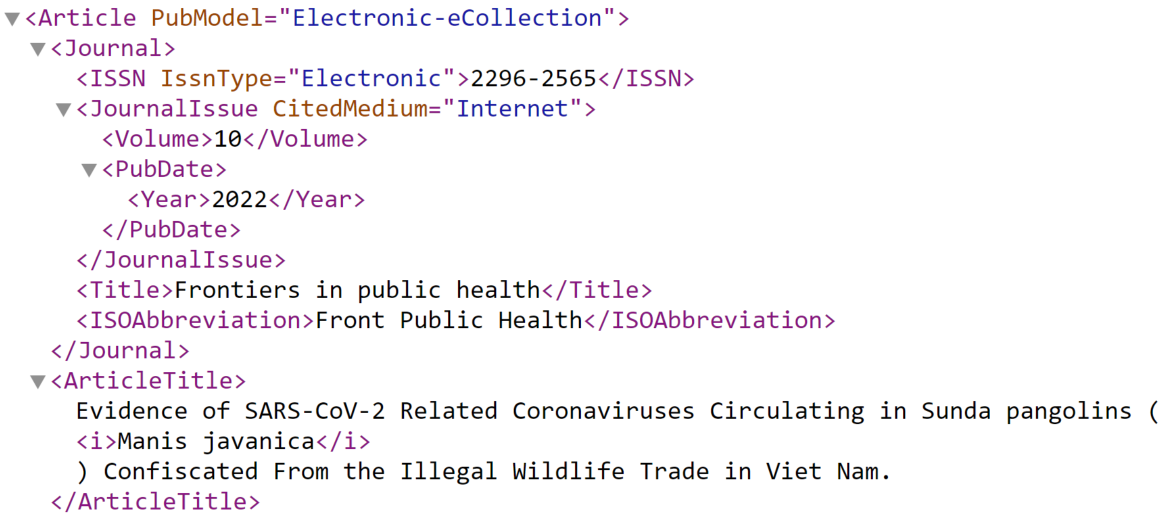
2) development of biodiversity-specific annotation workflows to deliver powerful text analytics in the domain of biotic interactions,

The graphic user interface is shown in the next section (Results) together with the evaluation of the services.

2.1. Data description and acquisition

MEDLINE & PubMed Central (PMC) are two complementary open literature resources maintained by the US National Library of Medicine (NLM). MEDLINE is a bibliographic database, i.e. it contains bibliographic information – such as title, authors, journal, abstract, along with some descriptors added by the NLM's indexers – of scientific articles published in a set of more than 5,000 high value biomedical journals. The collection contains more than 34M bibliographic references as of 2022. On the other hand, PMC is an archive of Open Access full texts. Moreover, PMC not only offers the published PDFs, but also a standardised and annotated version following the Journal Article Tag Suite (JATS) standard , as well as a large set of supplementary material files. PMC contains fewer articles than MEDLINE, around 4.7M in 2022, because many articles, recent or old, are not available under Open Access. Both collections are synchronised daily via the US National Library of Medicine (NLM) FTP servers and are made available after semantic enrichment in the SIB Literature Services (SIBiLS).

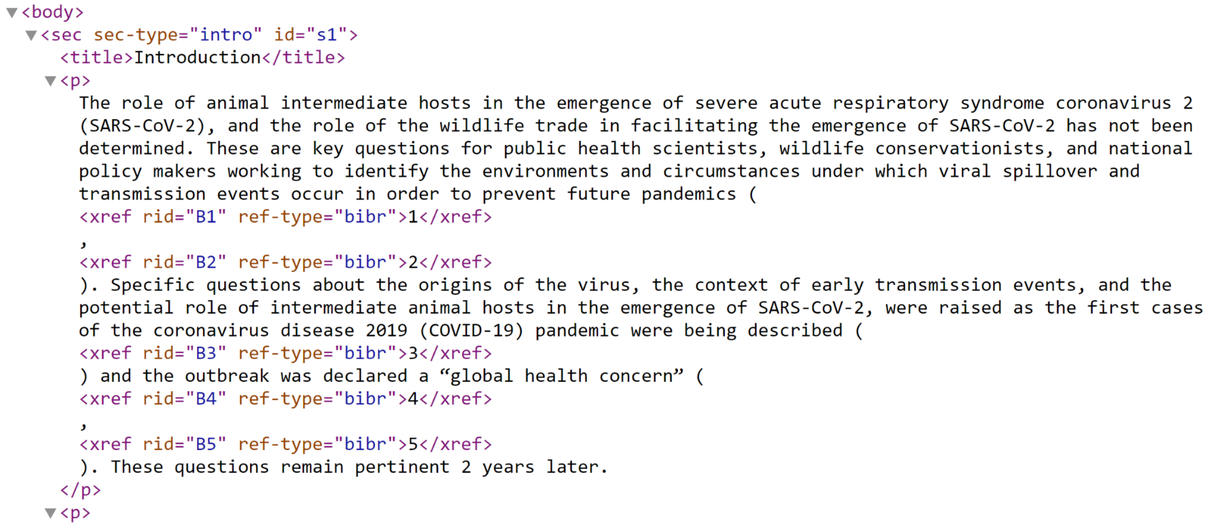
Here is an extract from a MEDLINE bibliographic record, fully accessible at https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=35356028&retmode=xml



```
▼<Article PubModel="Electronic-eCollection">
  ▼<Journal>
     <ISSN IssnType="Electronic">2296-2565</ISSN>
    ▼<JournalIssue CitedMedium="Internet">
       <Volume>10</Volume>
      ▼<PubDate>
         <Year>2022</Year>
       </PubDate>
     </JournalIssue>
     <Title>Frontiers in public health</Title>
     <ISOAbbreviation>Front Public Health</ISOAbbreviation>
   </Journal>
  ▼<ArticleTitle>
     Evidence of SARS-CoV-2 Related Coronaviruses Circulating in Sunda pangolins (
     <i>Manis javanica</i>
     ) Confiscated From the Illegal Wildlife Trade in Viet Nam.
   </ArticleTitle>
```

Figure 1: *Excerpt of a MEDLINE citation as provided by the NLM.*

Here is the corresponding full text in PubMed Central, fully accessible at https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pmc&id=8959545&retmode=xml



```
▼<body>
  ▼<sec sec-type="intro" id="s1">
     <title>Introduction</title>
    ▼<p>
       The role of animal intermediate hosts in the emergence of severe acute respiratory syndrome coronavirus 2
       (SARS-CoV-2), and the role of the wildlife trade in facilitating the emergence of SARS-CoV-2 has not been
       determined. These are key questions for public health scientists, wildlife conservationists, and national
       policy makers working to identify the environments and circumstances under which viral spillover and
       transmission events occur in order to prevent future pandemics (
       <xref rid="B1" ref-type="bibr">1</xref>
       ,
       <xref rid="B2" ref-type="bibr">2</xref>
       ). Specific questions about the origins of the virus, the context of early transmission events, and the
       potential role of intermediate animal hosts in the emergence of SARS-CoV-2, were raised as the first cases
       of the coronavirus disease 2019 (COVID-19) pandemic were being described (
       <xref rid="B3" ref-type="bibr">3</xref>
       ) and the outbreak was declared a “global health concern” (
       <xref rid="B4" ref-type="bibr">4</xref>
       ,
       <xref rid="B5" ref-type="bibr">5</xref>
       ). These questions remain pertinent 2 years later.
     </p>
    ▼<p>
```

Figure 2: *Excerpt of the corresponding PMC full text as provided by the NLM.*

Beyond full-texts in JATS-annotated XML, PMC also provides supplementary materials submitted by the authors but not published in the PDF: images, spreadsheets, videos etc. Although these files may contain entities not mentioned in the full text (in particular gene variants), to our knowledge they are not exploited and indexed by any search engine other than SIBiLS. For each article, textual passages are extracted from supplementary images thanks to a Optical Characters Recognition (OCR) pipeline based on the Tesseract open solution, and from spreadsheets thanks to locally developed script (Naderi et al. 2022). As these pipelines are time consuming, they are currently only applied to a subset of potentially relevant articles, i.e. articles that contain a combination of keywords such as “variant” or “polymorphism”. As a result, in addition to the PubMed Central index, SIBiLS proposes a search engine in the supplementary materials for around 750,000 publications (about 20% of all PMC), gathering about 4 millions files. A distribution of the file index is shown below:

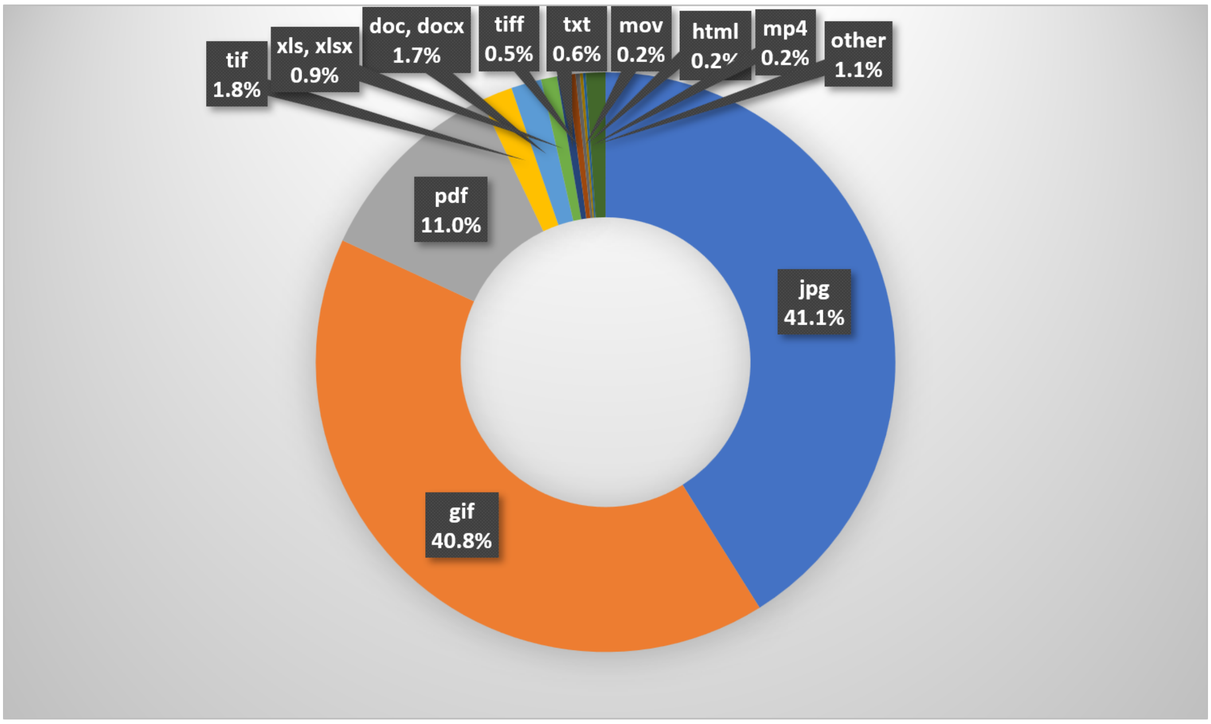
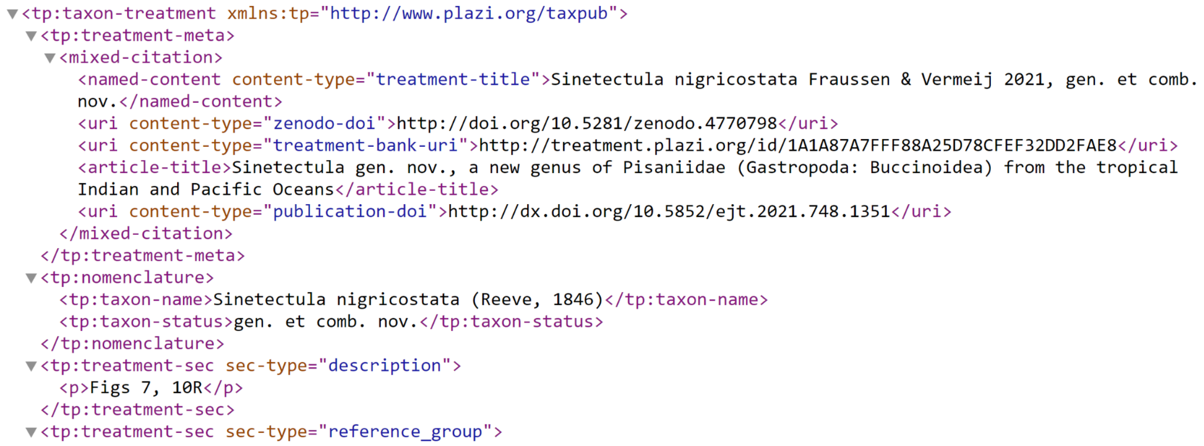


Figure 3: *Distribution of file extensions in the PMC supplementary Material.*

The search in supplementary data files including images is still in a beta version, however it is already in production settings within the Variomes variant-curation support application, see https://candy..ch/Variomes/.

Finally, TreatmentBank (TB) is a resource provided by Swiss Plazi GmbH to integrate taxonomic data from scholarly publications, often not in the scope of the previously described resources. Treatments are structured taxonomic descriptions, along with literature references, in JATS-XML format. In 2022, 450,000 treatments were duplicated by TreatmentBank onto a SIB FTP server, and from then on daily updated in SIBiLS.



```
▼<tp:taxon-treatment xmlns:tp="http://www.plazi.org/taxpub">
  ▼<tp:treatment-meta>
    ▼<mixed-citation>
       <named-content content-type="treatment-title">Sinetectula nigricostata Fraussen & Vermeij 2021, gen. et comb.
       nov.</named-content>
       <uri content-type="zenodo-doi">http://doi.org/10.5281/zenodo.4770798</uri>
       <uri content-type="treatment-bank-uri">http://treatment.plazi.org/id/1A1A87A7FFF88A25D78CFEF32DD2FAE8</uri>
       <article-title>Sinetectula gen. nov., a new genus of Pisaniidae (Gastropoda: Buccinoidea) from the tropical
       Indian and Pacific Oceans</article-title>
       <uri content-type="publication-doi">http://dx.doi.org/10.5852/ejt.2021.748.1351</uri>
     </mixed-citation>
   </tp:treatment-meta>
  ▼<tp:nomenclature>
     <tp:taxon-name>Sinetectula nigricostata (Reeve, 1846)</tp:taxon-name>
     <tp:taxon-status>gen. et comb. nov.</tp:taxon-status>
   </tp:nomenclature>
  ▼<tp:treatment-sec sec-type="description">
     <p>Figs 7, 10R</p>
   </tp:treatment-sec>
  ▼<tp:treatment-sec sec-type="reference_group">
```

Figure 4: *Excerpt of a Plazi treatment as provided by TreatmentBank.*

More library contents are continuously being added to SIBiLS. We are working with Pensoft to grow the coverage of biodiversity journals and the ClinicalTrials.gov reports are also candidates for integration into SIBiLS.

2.1.1. Use case

The D11.1 Evaluation benchmark: a database of questions associated with a set of relevant articles (e.g. PMIDs) to measure the progress of the WP towards an effective search service contains a list of factoid questions and two use cases. The following factoid questions are use cases of passage retrieval services:

- Biotic interaction
  - 1.1 What species are predators of bats?
  - 1.3 What diseases are transmitted by ticks?
  - 1.15 What are the transmission routes of coronavirus?
- Occurrences
  - 1.17 Give me all papers where human ABL1(F359V) is mentioned in full-text tables
  - 1.18 Give me all papers where human ABL1(F359V) is mentioned in supplementary data images
  - 1.19 Give me all papers where A.Y.42 variant of SARS-Cov-2 is mentioned in conclusion
  - 1.22 What species entities occur in a given paper/abstract/section?

Although ultimately, D11.4 will provide an answer to a question, D11.3 is an interim product as it only tries to identify a passage - i.e. a relatively short snippet of text with maximally includes a couple of sentences - where the answer could then be inferred.

In addition, the passage retrieval services can provide answers for some of the steps of the second use case, "Find which species interact with the sequence owner (*) and is involved in Myiasis" (see D11.1 to get the sequence). The complete workflows to implement D11.1 use cases is shown in Figures 5 and 6 below:

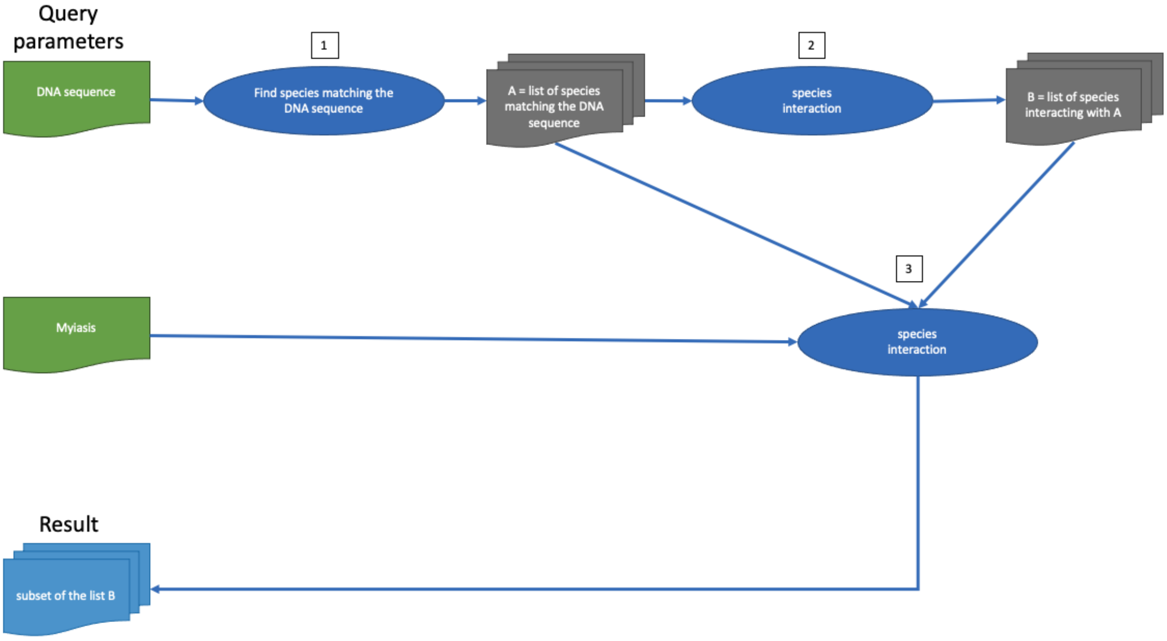

Query parameters
DNA sequence
1
Find species matching the DNA sequence
A = list of species matching the DNA sequence
2
species interaction
B = list of species interacting with A
3
Myiasis
species interaction
Result
subset of the list B

Figure 5: *Workflows describing how sequence specific mutations (i.e. Single Nucleotide Polymorphisms or SNPs) and biotic interactions can be combined to answer complex user queries.*

And step by step:

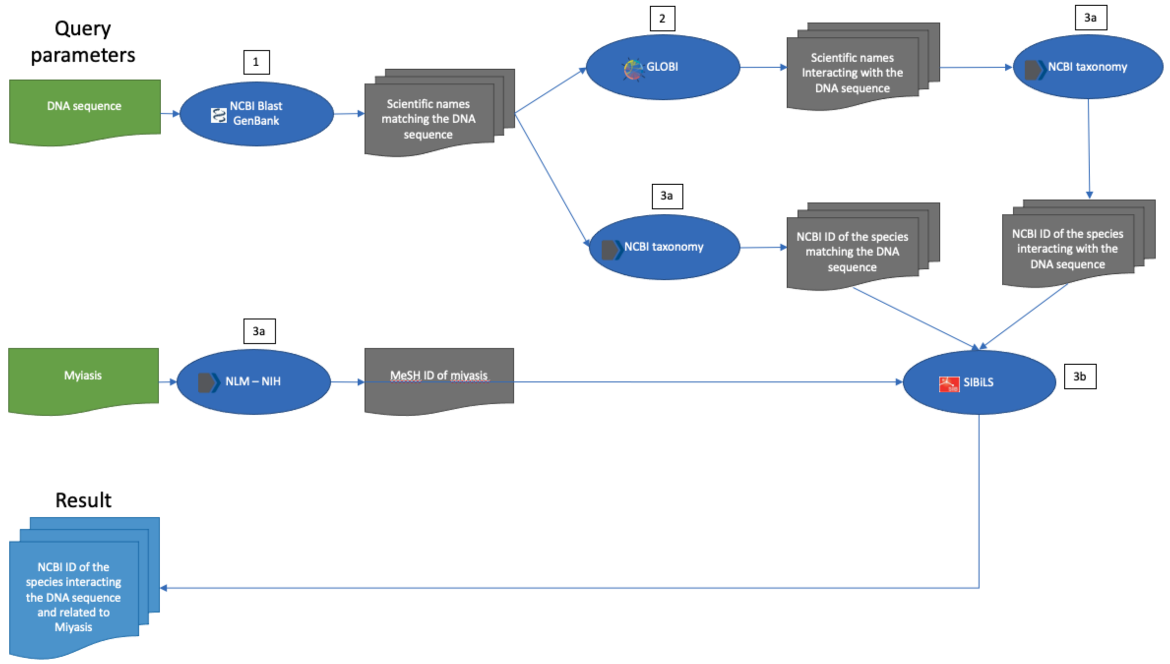

Query parameters
1
DNA sequence
NCBI Blast GenBank
Scientific names matching the DNA sequence
2
GLOBI
Scientific names Interacting with the DNA sequence
3a
NCBI taxonomy
3a
NCBI taxonomy
NCBI ID of the species matching the DNA sequence
NCBI ID of the species interacting with the DNA sequence
3a
Mylasis
NLM – NIH
MeSH ID of miyasis
SIBiLS
3b
Result
NCBI ID of the species interacting the DNA sequence and related to Miyasis

Figure 6: *Step wise combination of sequence specific mutations (SNPs) and biotic interactions. Biotic interactions can either explore tabular data, using GLOBI, or the literature. The NCBI taxonomy was used in D11.1 to build the use cases, while SIBiLS can accommodate several taxonomic backbones, including the Open Tree of Life (OToL), which includes and expands the NCBI Taxonomy, and includes also CoL synonyms.*

In step 2, GLOBI provides a list of species interacting with Dermatobia hominis (the answer in step 1). Alternatively, the passage retrieval services provide the answer to this query.

In step 3b, we display the JSON results from the REST API. This document presents an interface allowing the user to interact with the result.

2.1.2. Data volume and sources

Here is a synthetic view of the data volumes in SIBiLS :

	MEDLINE	PubMed Central	Supp. data	Plazi treatments
Volume in 2022	34.6M	4.7M	750K	450K
Growth in 2021	+ 1.3M	+ 220K	+ 30K	+ 30K

Table 1: *Volume and annual growth for collections included in SIBiLS.*

2.2. Architecture of services

In this section, we describe how the new services, which are built on top of SIBiLS and TreatmentBank, are integrated.

2.2.1. SIBiLS workflow

The next figure illustrates the general workflow of SIBiLS (Gobeill et al. 2020).

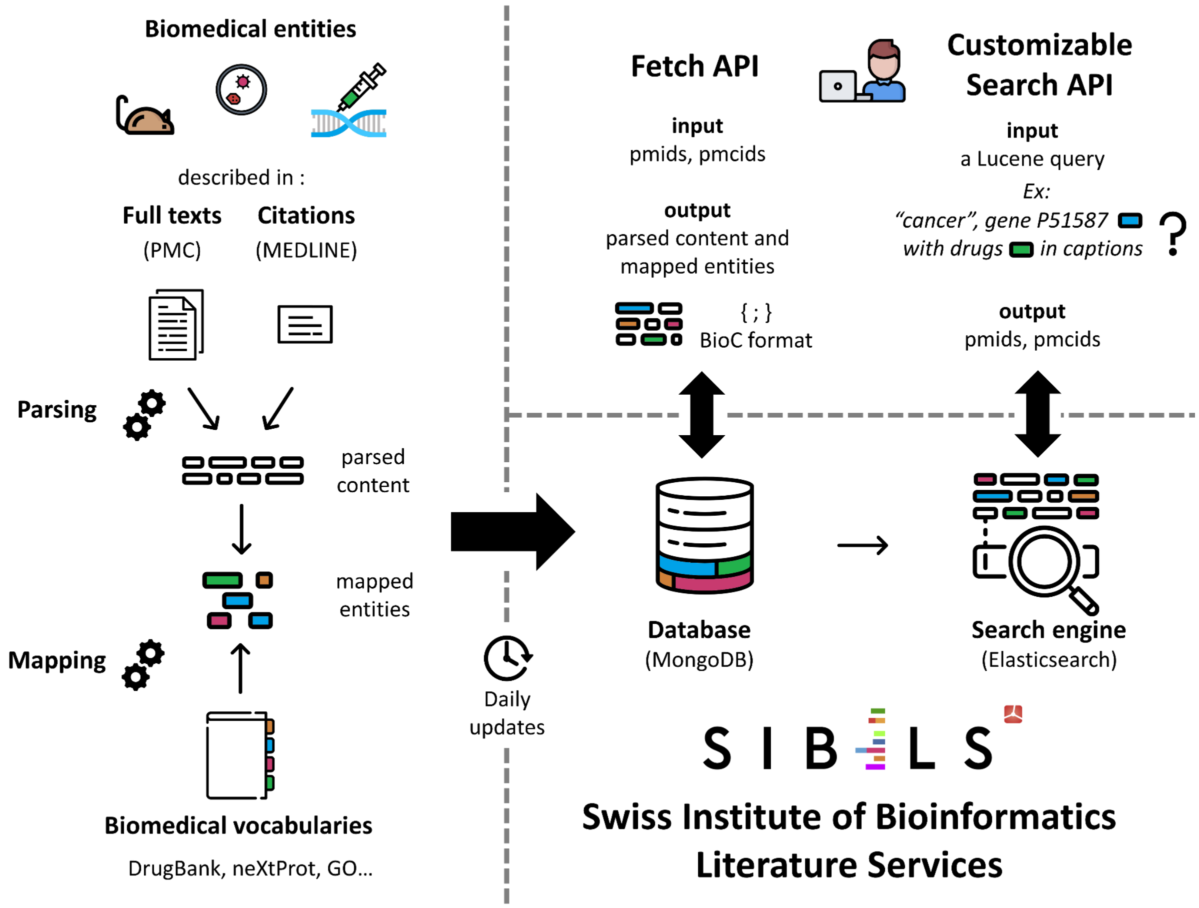


Figure 7: *Data workflows in SIBiLS. PMC and MEDLINE are harvested, parsed and annotated with the vocabulary collections (left panel). The APIs are described on the right panel.*

The Fetch APIs allow the user to retrieve annotated contents from MEDLINE or PMC. The input is a set of pmids, or pmcids (up to 1000 per request). The output is a set of parsed and annotated contents, in both JATS and BioC formats.

The Customizable Search APIs allow the user to perform a fully customizable search for valuable documents in MEDLINE or PMC Open Access. The power of these services is based on the efficiency of Elasticsearch engines, and on the rich Lucene query language, which allows to investigate a large panel of searching strategies.

2.2.2. Harvesting of NLM and Plazi's contents

Collections are acquired from providers (such as the NLM or Plazi), and updated daily. For MEDLINE & PMC, new, updated or deleted records are provided everyday in a dump file and made available in a FTP server. For Plazi, TreatmentBank updates treatments in real time in a dedicated SIBiLS FTP server. The SIBiLS pipeline uses local XML parsers. Their main role is to generate a simple JSON record, representing the different fields specific to the document. The json document representations are stored in a MongoDB database, ready to be accessed by the automatic annotation tool and the search engine. The SIBiLS parser turns the hierarchical structure of the document into a flat list of sections, each of which acting as a container for a flat list of multiple contents. The final representation is simple, easy to process and reflects the original sequential position of text elements as well as their hierarchical level in the document structure.

2.2.3. Automatic annotation pipeline

The annotation pipeline is divided in four steps. (i) Extraction: for a given document, the parsed representation is loaded from the MongoDB database, and fields of interest are extracted. Some are common to both collections (title, abstract, keywords), while others are more specific like MeSH terms for MEDLINE citations, or elements relative to figures or tables for full texts. (ii) Tokenization: each sentence is broken down into individual words and words n-grams (sequences of words). (iii) String pre-processing: it consists in dealing with special characters. For example, words containing a dash are transformed to a set of additional words ('B-RAF' becomes 'B', 'RAF', 'BRAF'), and symbols are replaced by corresponding Latin alphabet letters ('β' becomes 'b'). (iv) Annotations: they are produced thanks to lexical mapping between the pre-processed strings and the exploited vocabularies. For each vocabulary concept, the set of possible strings (e.g. preferred term, synonym) is tentatively matched in the text, and eventually results in annotations.

2.2.4. Indexing

The content parsing and automatic annotation pipelines deliver up-to-date json representations in a MongoDB database. Then, both representations are combined and indexed in Lucene Elasticsearch search engines. A Tomcat Web server handles requests from API clients. For the fetch APIs, parsed content and their annotations are beforehand converted and stored in BioC format, allowing the API to return the requested data in optimised response times. The maximum number of documents that can be requested per call is 1000. The search APIs submit a Lucene query to the Elasticsearch engines, and return the engine result set in its native json format.

2.3. Annotation workflows

The biodiversity-specific annotations workflow relies on the vocabulary-based annotations of the literature, as provided by SIBiLS. In addition to the wealth of named entities already annotated by SIBiLS, two new terminologies are used: the Open Tree of Life to support the recognition of species and ROBI for recognizing biotic interactions. Open Tree of Life is a combination of several large classifications and thus provides a fairly comprehensive vocabulary for species. It consists of more than 4,5 millions of unique concepts and encompasses more than 6,7 millions of terms. The ROBI terminology is a small vocabulary, which has been manually extended - also thanks to the contribution of the CETAF/DiSSCo working group (Poelen et al. 2022) working on virus-related biotic interactions. In our extended version, it contains 64 concepts and 73 terms.

Our approach is based on two steps. First, we pre-process the collections to create triplets corresponding to an interaction. For instance, in the sentence “Biomphalaria glabrata is a major intermediate host for the helminth parasite Schistosoma mansoni.”, we extract an interaction between “Biomphalaria glabrata” and “Schistosoma mansoni” with an interaction of “host of”. Second, we build a search engine able to search in the pre-processed interactions database. Our aim is to build triplets mentioning two species and an interaction in the same passage.
The pre-processing to build triplets is based on four steps (Figure 8):

- Document split in passages. In this first version, we have defined the passage as a sentence. Each document from our collections (i.e. MEDLINE, Plazi and PMC) are processed.
- Extraction of all annotations of a passage from the SIBiLS MongoDB database. Only annotations of species (Open Tree of Life) and biotic interactions (ROBI) are taken into account.
- Annotation processing. Annotations are processed to retrieve the annotation type, the concept string and its position in the passage.
- Triplets building. All possible interactions are built as triplets. If more than two species are identified in the passage, several triplets are built, each mentioning a different couple of interacting species. When two identified species are overlapping in the text, no interaction is built for the given couple. If no biotic interaction term has been identified in the passage, the two species are stored without a specified interaction.
- The triplet annotations are then stored in a MongoDB database for further use by the search engine. For each triplet, the passage and the document are also stored.

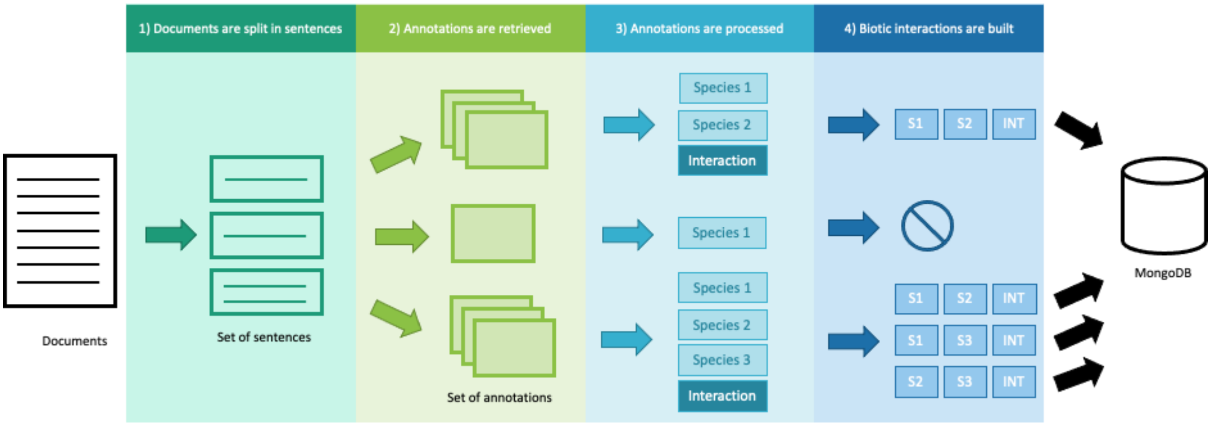


Figure 8: *Workflow to build the benchmark, i.e. the triplets of two species involved in a particular ROBI interaction.*

The search engine enables searching in the MongoDB collection. The search is based on two steps:

- Query expansion. The query expansion consists of expanding the query based on the tree hierarchy of the Open Tree of Life taxonomy. For instance, if a user query for interactions with Manis, we search for interaction of Manis, but also of Manis javanica, Manis gigantea, etc.

- Results ranking. The triplets are then ranked in order to return first triplets having the larger set of supporting passages and documents.

2.3.1. Existing annotations

Table 2 provides an overview of today's (from Oct 2022) annotations in SIBiLS, i.e. before the semantic enrichment resulting from the annotation of biotic interactions. Annotations in this table are limited to onto-terminological annotations, i.e. annotations resulting from an entity found in a terminology or an ontology. Other types of annotations, such as the accession numbers (e.g., Cellosaurus, UniProt or Plazi) as well as the binary of ternary relationships (e.g., between a gene or gene product and a sequence) are not listed here.

Terminology		Number of annotations	
		Medline (#docs: 34,676,223)	PMC (#docs: 4,734,800)
Affiliations		13,185,097	4,788,274
ATC		92,988,510	101,397,006
Cellosaurus		100,905,348	346,515,102
Chebi		261,761,748	343,295,442
COVOC	Biological/Medical Vocabulary	121,576,194	229,600,350
	Conceptual Entities	56,215,638	126,705,504
	Chemicals	1,991,011	4,399,924
	Cell Lines	5,669	120,275
	Clinical Trials	15	2,324
	Diseases and Syndromes	19,440,679	31,441,106
	Geographic locations	3,461,011	6,269,278
	Organisms	9,197,528	24,581,035
	Proteins and Genomes	3,249,731	8,549,832
DisProt	Type 1	3,707,482	7,836,612
	Type 2	7,770,379	13,748,622
	Type 3	9,395,191	14,894,900
	Type 4	11,284,944	15,769,382
Drugbank		141,914,927	171,378,846

ECO		2,592,290	8,220,457
ENVO		11,123,739	28,797,444
GO	Biological Process	43,235,240	67,149,137
	Cellular Component	12,007,186	22,637,716
	Molecular Function	18,215,310	26,015,659
ICDO3		7,619,421	6,833,304
MESH		1,107,070,232	1,048,439,766
Metadata		5,724,692	7,652,559
NCBI	Clinical	24,103,860	35,786,858
	Full	247,690,726	93,027,869
NCI Thesaurus		377,433,452	536,267,898
neXtProt		86,889,001	266,913,758
Open Tree of Life		114,949,839	221,034,294
PPI PTM		4,197,405	11,375,848
ROBI		295,691	911,370
UniProt_sprot		2,765,809,465	176,065,115
	Total	5,687,008,651	4,008,422,866

Table 2: *Statistical distribution of onto-terminological annotations within the SIB Literature Services before enrichment with entities relevant to identify biotic interactions.*

With nearly 9.7 billion annotations, a total exceeding 10 billion is expected once OToL species and ROBI will be annotated.

2.3.2. Species annotations

Here we report specifically on statistical distributions (Table 3) related to terminologies used to build the biotic interactions triplets: Open Tree of Life and ROBI.

Terminology	**Vocabulary size**	**Number of annotations**
Open Tree of Life	4,528,126 concepts and 6,753,382 terms	Plazi: 11,285,162
		Medline: 114,949,839
		PMC: 221,034,294

ROBI	64 concepts and 73 terms	Plazi: 4,553
		Medline: 295,691
		PMC: 911,370

Table 3: *Statistical distributions of SIBiLS annotations needed to build the biotic interactions' triplets*

2.3.3. Biotic interactions

Here we report on statistics (Table 4) about the biotic interaction triplets built based on two collections: MEDLINE and Plazi. We distinguish between "complete interactions", which are defined as the occurrence of two different species and one biotic ROBI interaction found within a given sentence, and "species co-occurrences", which are simply defined as two different species found in a given sentence.

	Plazi	**Medline**
Complete interactions with ROBI	78,999	106,025
Species co-occurrences (i.e. two species in the same sentence without ROBI concept)	20,079,603	39,116,455
Total number of interactions	20,158,602	39,222,480
Number of unique interactions	6,761,297	7,527,731

Table 4: *Statistics for the biotic interaction triplets for MEDLINE and Plazi.*

2.3.4. Quality control

From Table 4, we observe that the MEDLINE and Plazi collections show relatively comparable volumes of biotic interactions. By combining the two sources of information, SIBiLS is likely the largest source of evidence to explore biotic interactions.

However, the number of false positives is relatively high - about 30% based on a manually controlled sample (N=100). Data cleaning processes will therefore be necessary before releasing the new library collection. This includes applying frequency filters to remove highly frequent general English words, which are lexically ambiguous and which happen to be also species names (e.g. "here", "data"). Similarly species names, which are also anatomical or location entities (e.g. "pes", "patella", "argentina", "china") will have to be filtered out. Several parallel strategies are being considered to deliver a more robust annotated collections, including crowd curation services to help collect feedback from end-users.

3. Results

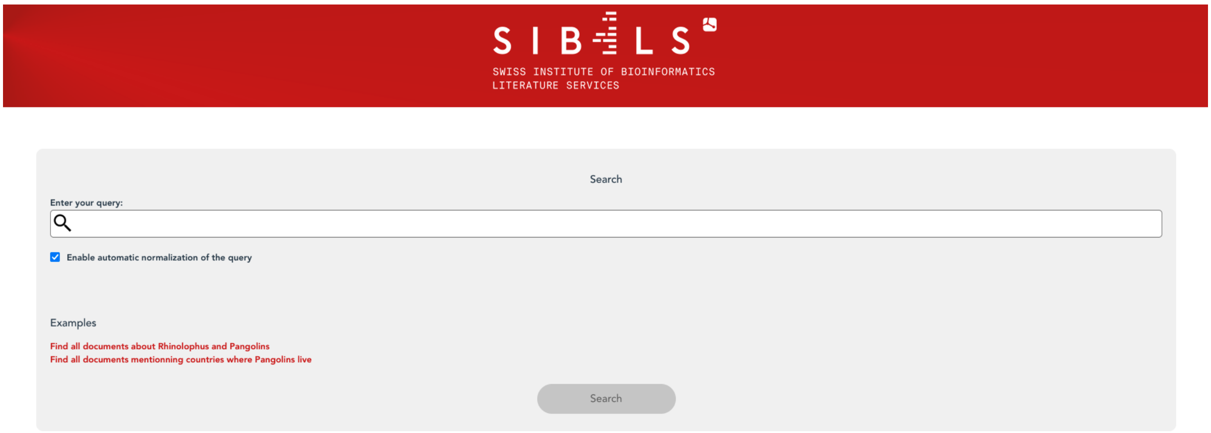
First, we introduce the draft Graphical User Interface, which we decided to develop as a proof of concept and for the sake of testing of the API. The messaging services to interact with the API are also shown. Further, we provide an evaluation of the robustness of our services based on response time. Finally, we evaluate the search effectiveness of the passage retrieval services.

3.1. User interface

Two user interfaces are proposed. The first one enables users to search for documents in the SIBiLS database. The second one enables users to search for biotic interactions in a sentence. Other window sizes, e.g. biotic interactions occurring in a given passage, could be considered in the future. In particular, the planned processing of biotic interactions within full-text articles from PubMed Central will demande the design of other indexing units.

3.1.1. SIBiLS User Interface

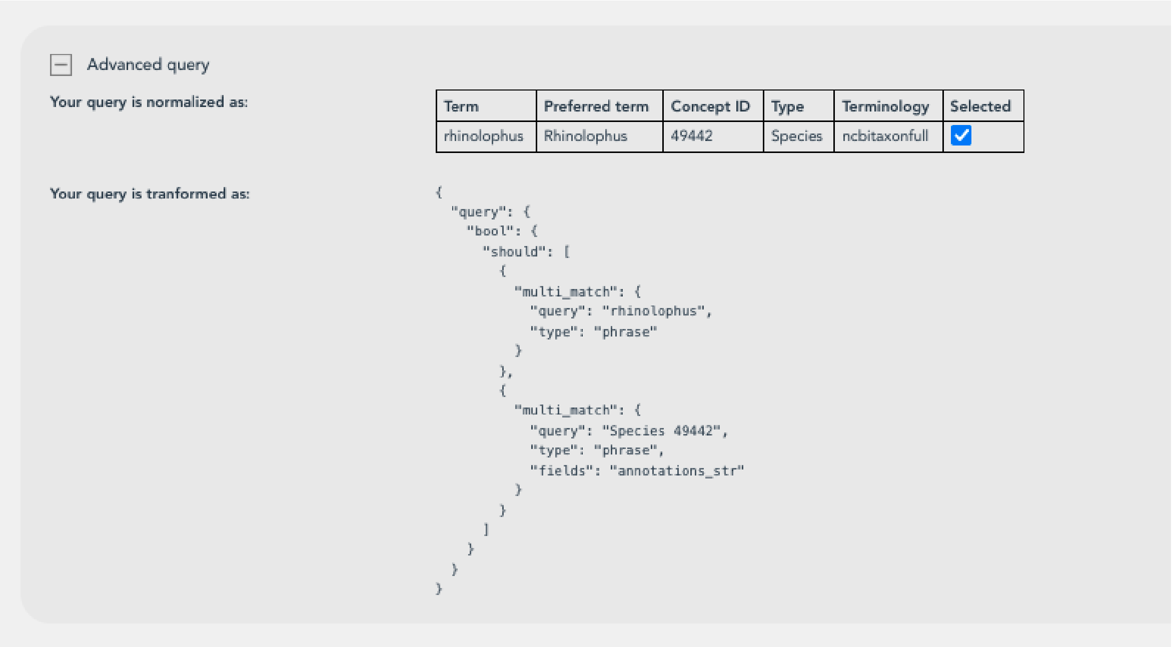
A user interface has been developed to visualise the documents retrieved in SIBiLS for Medline, PubmedCentral and Plazi. The landing page (Figure 9) enables the user to type a query. The query is based on boolean syntax (e.g. pangolin AND rhinolophus). The query is then automatically processed and the user can optionally review and edit this processing (Figure 10). The processing consists of normalising the query to map the text to concept identifiers from the SIBiLS terminologies, as well as converting the boolean query to a JSON query. The user can then search and the three collections are searched in parallel. Once results are ready, they are displayed in a table (Figure 11 and Figure 12). Facets are available to filter the results, based on tagged entities, such as species and on article metadata, such as publication type, journal, etc. Finally, the user can flag documents of interest and export them in JSON or CSV format (Figure 13).



☑ Enable automatic normalization of the query

Find all documents about Rhinolophus and Pangolins
Find all documents mentionning countries where Pangolins live

Figure 9: *Landing page of SIBiLS.*



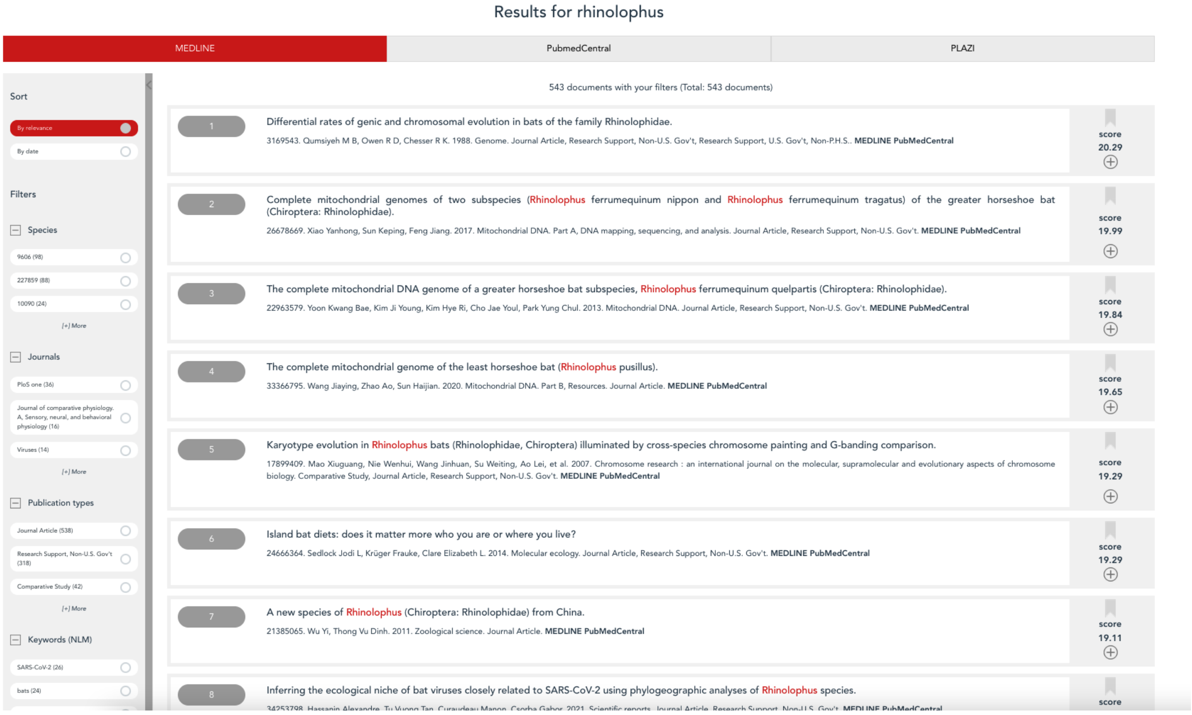
Your query is normalized as:

Term	Preferred term	Concept ID	Type	Terminology	Selected
rhinolophus	Rhinolophus	49442	Species	ncbitaxonfull	☑

Your query is tranformed as:

```
{
  "query": {
    "bool": {
      "should": [
        {
          "multi_match": {
            "query": "rhinolophus",
            "type": "phrase"
          }
        },
        {
          "multi_match": {
            "query": "Species 49442",
            "type": "phrase",
            "fields": "annotations_str"
          }
        }
      ]
    }
  }
}
```

Figure 10: *Example of a normalised query.*



Results for rhinolophus

543 documents with your filters (Total: 543 documents)

1 Differential rates of genic and chromosomal evolution in bats of the family Rhinolophidae.
3169543. Qumsiyeh M B, Owen R D, Chesser R K. 1988. Genome. Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, Non-P.H.S.. **MEDLINE PubMedCentral**
score 20.29 ⊕

2 Complete mitochondrial genomes of two subspecies (Rhinolophus ferrumequinum nippon and Rhinolophus ferrumequinum tragatus) of the greater horseshoe bat (Chiroptera: Rhinolophidae).
26678669. Xiao Yanhong, Sun Keping, Feng Jiang. 2017. Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis. Journal Article, Research Support, Non-U.S. Gov't. **MEDLINE PubMedCentral**
score 19.99 ⊕

3 The complete mitochondrial DNA genome of a greater horseshoe bat subspecies, Rhinolophus ferrumequinum quelpartis (Chiroptera: Rhinolophidae).
22963579. Yoon Kwang Bae, Kim Ji Young, Kim Hye Ri, Cho Jae Youl, Park Yung Chul. 2013. Mitochondrial DNA. Journal Article, Research Support, Non-U.S. Gov't. **MEDLINE PubMedCentral**
score 19.84 ⊕

4 The complete mitochondrial genome of the least horseshoe bat (Rhinolophus pusillus).
33366795. Wang Jiaying, Zhao Ao, Sun Haijian. 2020. Mitochondrial DNA. Part B, Resources. Journal Article. **MEDLINE PubMedCentral**
score 19.65 ⊕

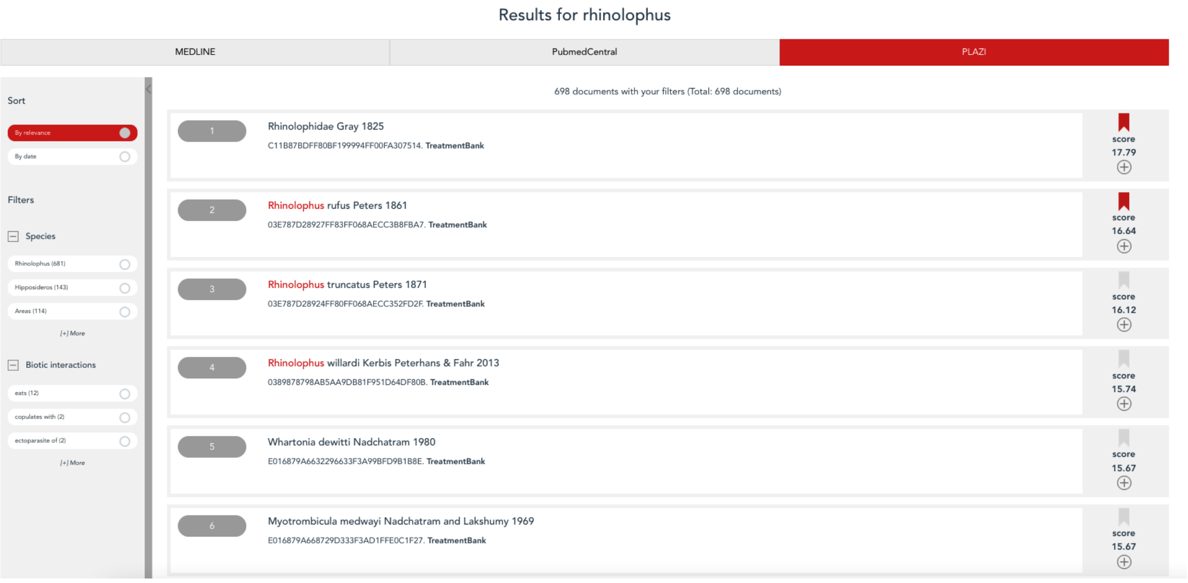
5 Karyotype evolution in Rhinolophus bats (Rhinolophidae, Chiroptera) illuminated by cross-species chromosome painting and G-banding comparison.
17899409. Mao Xiuguang, Nie Wenhui, Wang Jinhuan, Su Weiting, Ao Lei, et al. 2007. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology. Comparative Study, Journal Article, Research Support, Non-U.S. Gov't. **MEDLINE PubMedCentral**
score 19.29 ⊕

6 Island bat diets: does it matter more who you are or where you live?
24666364. Sedlock Jodi L, Krüger Frauke, Clare Elizabeth L. 2014. Molecular ecology. Journal Article, Research Support, Non-U.S. Gov't. **MEDLINE PubMedCentral**
score 19.29 ⊕

7 A new species of Rhinolophus (Chiroptera: Rhinolophidae) from China.
21385065. Wu Yi, Thong Vu Dinh. 2011. Zoological science. Journal Article. **MEDLINE PubMedCentral**
score 19.11 ⊕

8 Inferring the ecological niche of bat viruses closely related to SARS-CoV-2 using phylogeographic analyses of Rhinolophus species.
34253798. Hassanin Alexandre, Tu Vuong Tan, Curaudeau Manon, Csorba Gabor. 2021. Scientific reports. Journal Article, Research Support, Non-U.S. Gov't. **MEDLINE PubMedCentral**
score

Figure 11: *Display of MEDLINE publications retrieved in SIBiLS.*



Results for rhinolophus

698 documents with your filters (Total: 698 documents)

1 Rhinolophidae Gray 1825
C11B87BDFF80BF199994FF00FA307514. **TreatmentBank**
score 17.79

2 Rhinolophus rufus Peters 1861
03E787D28927FF83FF068AECC3B8FBA7. **TreatmentBank**
score 16.64

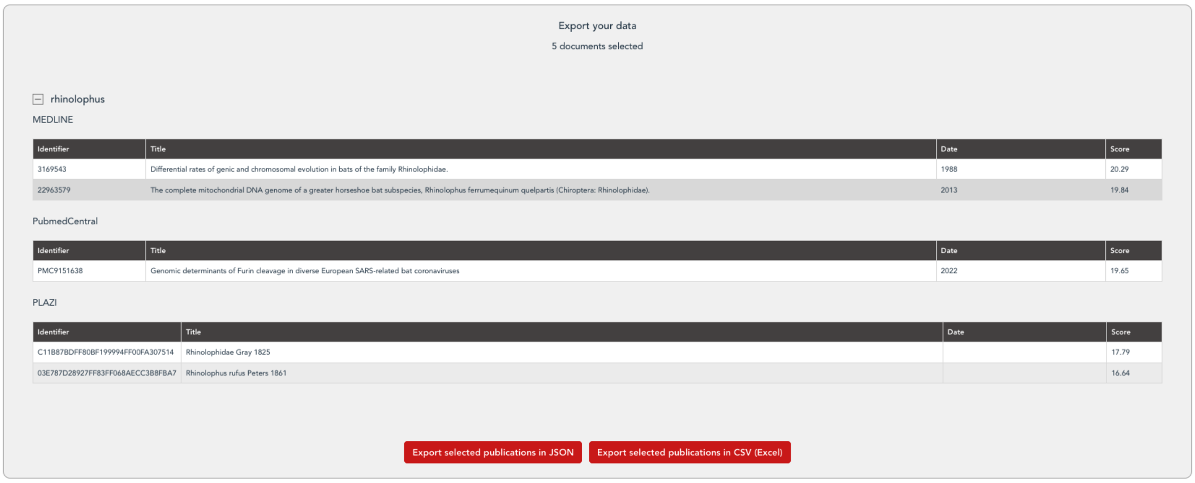
3 Rhinolophus truncatus Peters 1871
03E787D28924FF80FF068AECC352FD2F. **TreatmentBank**
score 16.12

4 Rhinolophus willardi Kerbis Peterhans & Fahr 2013
0389878798AB5AA9DB81F951D64DF80B. **TreatmentBank**
score 15.74

5 Whartonia dewitti Nadchatram 1980
E016879A6632296633F3A99BFD9B1B8E. **TreatmentBank**
score 15.67

6 Myotrombicula medwayi Nadchatram and Lakshumy 1969
E016879A668729D333F3AD1FFE0C1F27. **TreatmentBank**
score 15.67

Figure 12: *Display of Plazi publications retrieved in SIBiLS.*



Export your data

5 documents selected

rhinolophus

MEDLINE

Identifier	Title	Date	Score
3169543	Differential rates of genic and chromosomal evolution in bats of the family Rhinolophidae.	1988	20.29
22963579	The complete mitochondrial DNA genome of a greater horseshoe bat subspecies, Rhinolophus ferrumequinum quelpartis (Chiroptera: Rhinolophidae).	2013	19.84

PubmedCentral

Identifier	Title	Date	Score
PMC9151638	Genomic determinants of Furin cleavage in diverse European SARS-related bat coronaviruses	2022	19.65

PLAZI

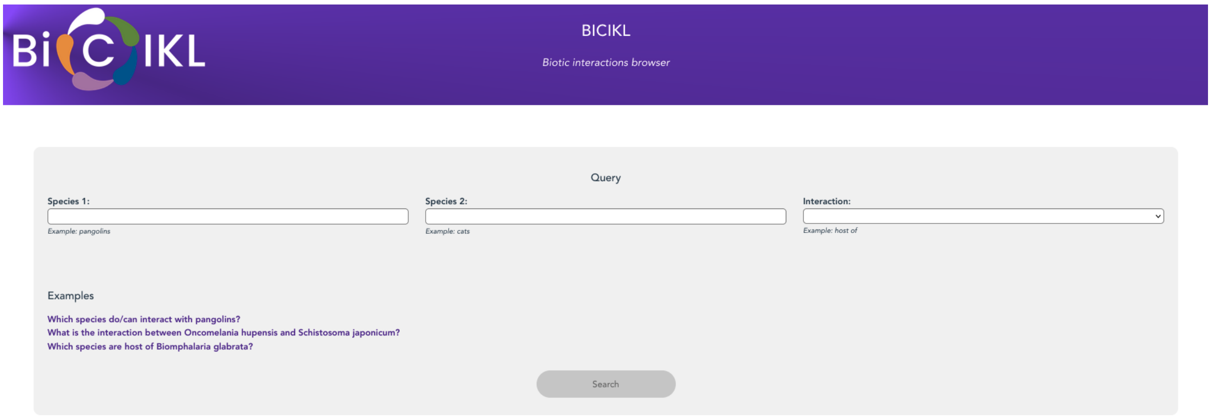
Identifier	Title	Date	Score
C11B87BDFF80BF199994FF00FA307514	Rhinolophidae Gray 1825		17.79
03E787D28927FF83FF068AECC3B8FBA7	Rhinolophus rufus Peters 1861		16.64

Export selected publications in JSON
Export selected publications in CSV (Excel)

Figure 13: *Export of documents of interest in SIBiLS.*

3.1.2. Biotic interaction triplets User Interface

A user interface has been developed to visualise the biotic interactions extracted from SIBiLS. The landing page (Figure 14) enables the end-users to search for one or two species as well as to select a biotic interaction from a list. The user can search either for a specific triplet to look at passages mentioning this triplet or search for any triplet involving one or two species. The species typed by the user are automatically normalised and the corresponding OpenTreeOfLife concept is suggested to the user (Figure 15). If alternative concepts have been found, a radiobutton list is proposed so that the user can select another concept if needed. The results (Figure 16) are splitted in tabs - one for each collection - and triplets are displayed in a table. For each triplet, the user can see the documents and passages (i.e. the sentences) containing the biotic interaction. In addition, facets are available to filter the results, based on entities and ROBI interactions identified.



BICIKL

Biotic interactions browser

Query

Species 1:

Example: pangolins

Species 2:

Example: cats

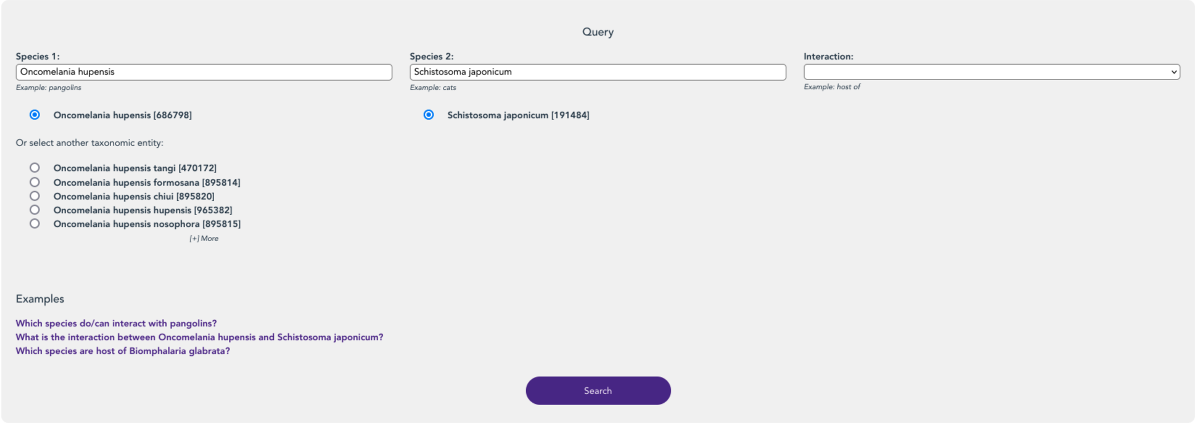
Interaction:

Example: host of

Examples

Which species do/can interact with pangolins?
What is the interaction between Oncomelania hupensis and Schistosoma japonicum?
Which species are host of Biomphalaria glabrata?

Figure 14: *Landing page of the biotic interaction service. It is available here: http://denver.hesge.ch/biotic/demo/ .*



Query

Species 1: Oncomelania hupensis

Example: pangolins

Species 2: Schistosoma japonicum

Example: cats

Interaction:

Example: host of

● **Oncomelania hupensis [686798]**

● **Schistosoma japonicum [191484]**

Or select another taxonomic entity:

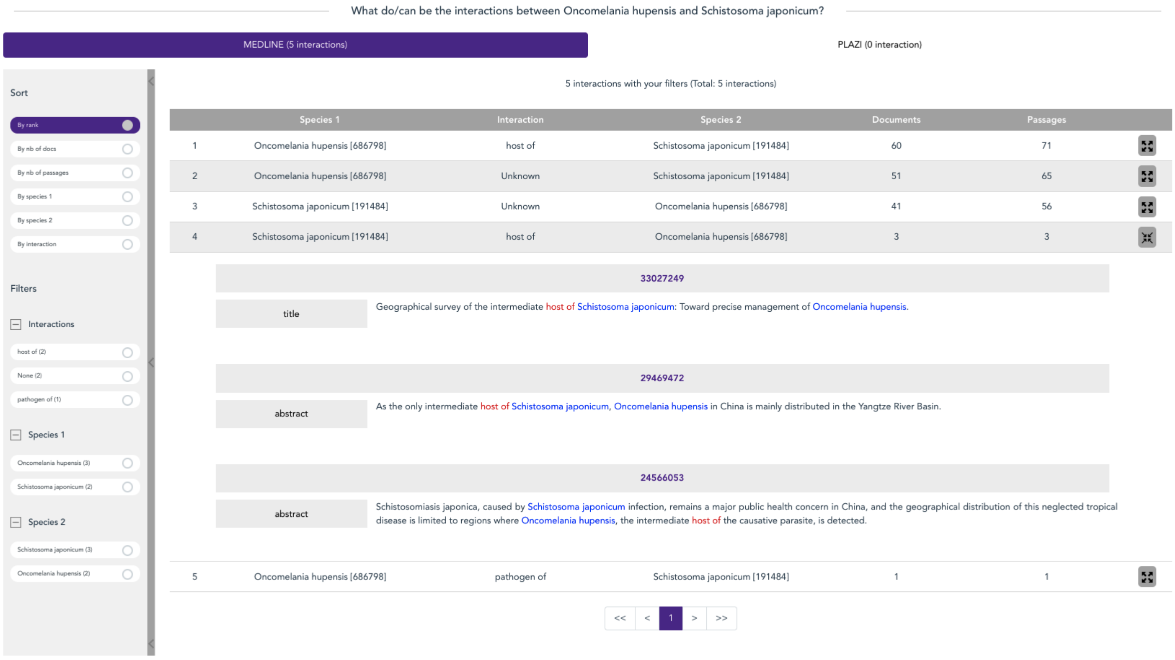
○ **Oncomelania hupensis tangi [470172]**
○ **Oncomelania hupensis formosana [895814]**
○ **Oncomelania hupensis chiui [895820]**
○ **Oncomelania hupensis hupensis [965382]**
○ **Oncomelania hupensis nosophora [895815]**

[+] More

Examples

Which species do/can interact with pangolins?
What is the interaction between Oncomelania hupensis and Schistosoma japonicum?
Which species are host of Biomphalaria glabrata?

Figure 15: *Example of a normalised query.*



What do/can be the interactions between Oncomelania hupensis and Schistosoma japonicum?

5 interactions with your filters (Total: 5 interactions)

	Species 1	Interaction	Species 2	Documents	Passages
1	Oncomelania hupensis [686798]	host of	Schistosoma japonicum [191484]	60	71
2	Oncomelania hupensis [686798]	Unknown	Schistosoma japonicum [191484]	51	65
3	Schistosoma japonicum [191484]	Unknown	Oncomelania hupensis [686798]	41	56
4	Schistosoma japonicum [191484]	host of	Oncomelania hupensis [686798]	3	3

33027249

title	Geographical survey of the intermediate host of Schistosoma japonicum: Toward precise management of Oncomelania hupensis.

29469472

abstract	As the only intermediate host of Schistosoma japonicum, Oncomelania hupensis in China is mainly distributed in the Yangtze River Basin.

24566053

abstract	Schistosomiasis japonica, caused by Schistosoma japonicum infection, remains a major public health concern in China, and the geographical distribution of this neglected tropical disease is limited to regions where Oncomelania hupensis, the intermediate host of the causative parasite, is detected.

5	Oncomelania hupensis [686798]	pathogen of	Schistosoma japonicum [191484]	1	1

Figure 16: *Biotic interactions retrieved by the service.*

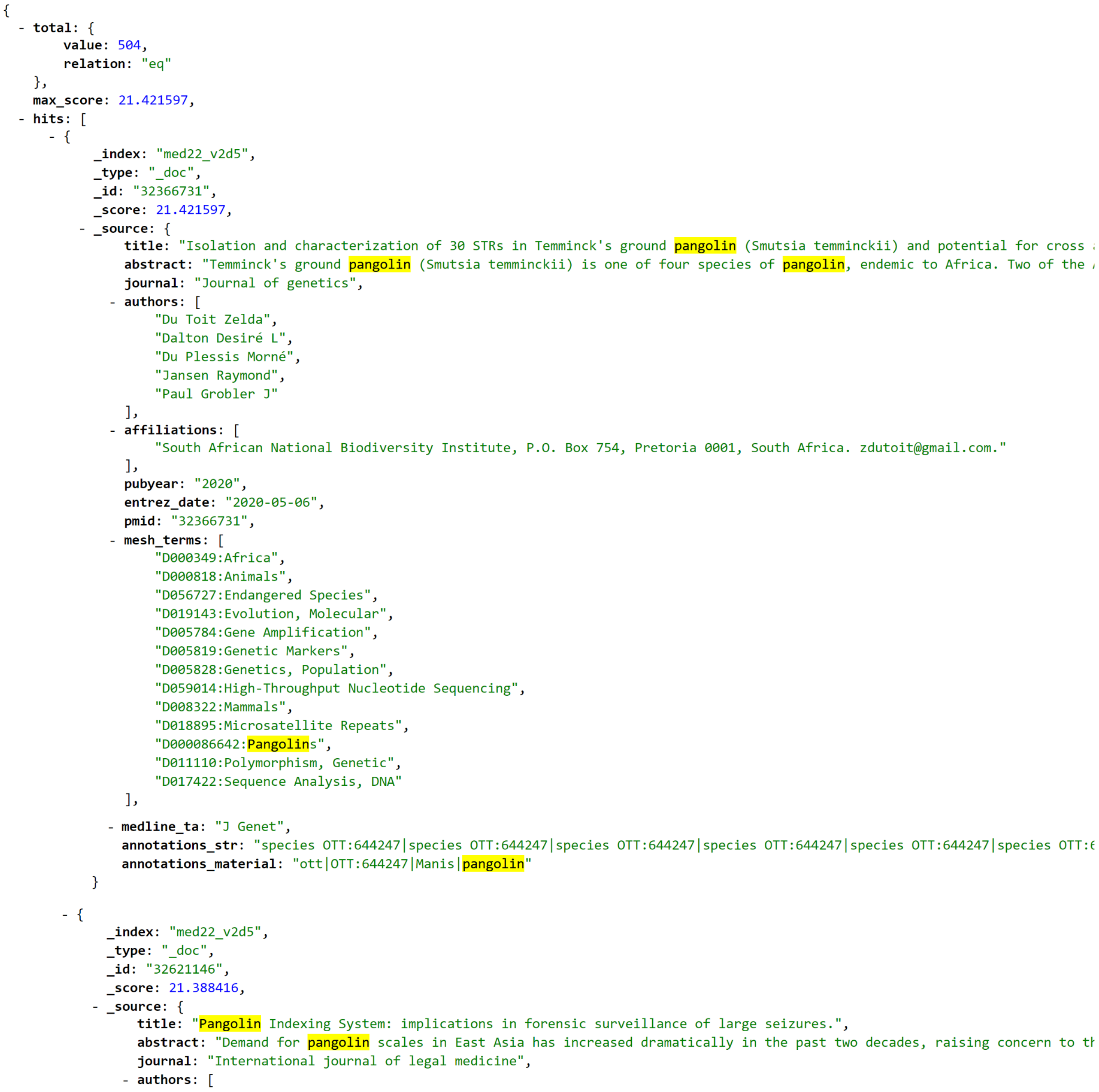
3.2. API messages

In this section, we present the API messages returned for both the SIBiLS APIs and biotic interaction triplets API.

3.2.1. SIBiLS APIs

The following figure illustrates a message returned by the SIBiLS search API in MEDLINE, with the query “Pangolin”. In this example, the first returned article contains several mentions of pangolin in the title, abstract, and MeSH terms. “Pangolin” was also identified by SIBiLS as an Open Tree of Life concept OTT:644247 Manis. The annotations_str field allows to exploit SIBiLS annotations and to retrieve documents dealing with the concept of pangolin, no matter if “Manis” or “Pangolin” was used by authors. Finally, the field annotation_material provides the GUI with all forms mapped in the text during annotations but is not indexed nor searchable.

The complete message can be accessed via the API:
https://candy.hesge.ch/SIBiLS/MEDLINE/v2.5/search.jsp?keywords=pangolin



```
{
  - total: {
        value: 504,
        relation: "eq"
    },
    max_score: 21.421597,
  - hits: [
      - {
            _index: "med22_v2d5",
            _type: "_doc",
            _id: "32366731",
            _score: 21.421597,
          - _source: {
                title: "Isolation and characterization of 30 STRs in Temminck's ground pangolin (Smutsia temminckii) and potential for cross
                abstract: "Temminck's ground pangolin (Smutsia temminckii) is one of four species of pangolin, endemic to Africa. Two of the
                journal: "Journal of genetics",
              - authors: [
                    "Du Toit Zelda",
                    "Dalton Desiré L",
                    "Du Plessis Morné",
                    "Jansen Raymond",
                    "Paul Grobler J"
                ],
              - affiliations: [
                    "South African National Biodiversity Institute, P.O. Box 754, Pretoria 0001, South Africa. zdutoit@gmail.com."
                ],
                pubyear: "2020",
                entrez_date: "2020-05-06",
                pmid: "32366731",
              - mesh_terms: [
                    "D000349:Africa",
                    "D000818:Animals",
                    "D056727:Endangered Species",
                    "D019143:Evolution, Molecular",
                    "D005784:Gene Amplification",
                    "D005819:Genetic Markers",
                    "D005828:Genetics, Population",
                    "D059014:High-Throughput Nucleotide Sequencing",
                    "D008322:Mammals",
                    "D018895:Microsatellite Repeats",
                    "D000086642:Pangolins",
                    "D011110:Polymorphism, Genetic",
                    "D017422:Sequence Analysis, DNA"
                ],

              - medline_ta: "J Genet",
                annotations_str: "species OTT:644247|species OTT:644247|species OTT:644247|species OTT:644247|species OTT:644247|species OTT:6
                annotations_material: "ott|OTT:644247|Manis|pangolin"
            }

        - {
              _index: "med22_v2d5",
              _type: "_doc",
              _id: "32621146",
              _score: 21.388416,
            - _source: {
                  title: "Pangolin Indexing System: implications in forensic surveillance of large seizures.",
                  abstract: "Demand for pangolin scales in East Asia has increased dramatically in the past two decades, raising concern to th
                  journal: "International journal of legal medicine",
                - authors: [
```

Figure 17: *Message returned by the dedicated SIBiLS search API in MEDLINE, with the query "Pangolin". The term was artificially highlighted in a Web browser.*

3.2.2. Biotic interaction triplets API

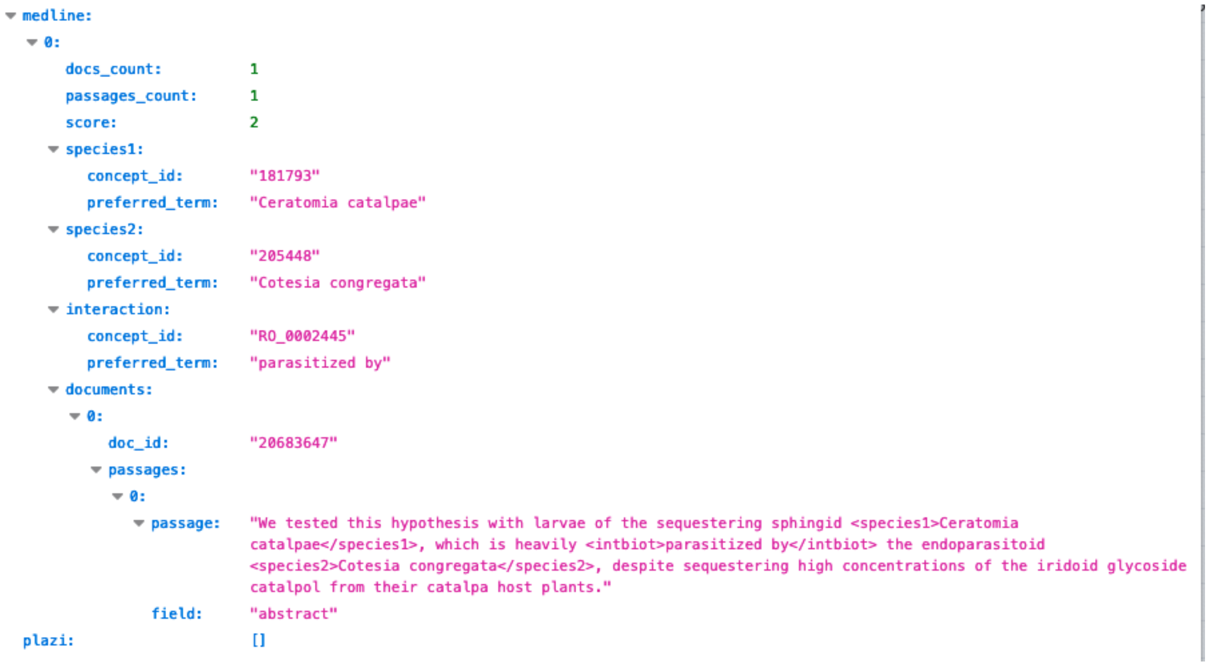
The biotic interaction triplets API (https://denver.hesge.ch/biotic/api/interactions) proposes three parameters (Table 5): a list of species, an interaction and a collection.

Parameter name	Type	Example
species	list of Open Tree of Life identifiers	181793, 205448
interaction	ROBI identifier	RO_0002445
collection	name of the collection to	medline

	search in	

Table 5: *Parameters for the biotic interaction triplets API.*

The output (Figure 18) of this API returns for each collection, a list of triplets. For each triplet, the following information is returned: the number of documents containing the triplet, the number of passages mentioning the triplet, a score (i.e. the sum of passages and documents), the two species involved (including the preferred term and its OpenTreeOfLife identifier), the biotic interaction identified (“None” if no interaction was retrieved) and the list of documents containing the triplet. For each document, a list of passages mentioning the triplets is proposed with tagged entities (i.e. species and biotic interaction).



```
▾ medline:
  ▾ 0:
      docs_count:          1
      passages_count:      1
      score:               2
    ▾ species1:
        concept_id:        "181793"
        preferred_term:    "Ceratomia catalpae"
    ▾ species2:
        concept_id:        "205448"
        preferred_term:    "Cotesia congregata"
    ▾ interaction:
        concept_id:        "RO_0002445"
        preferred_term:    "parasitized by"
    ▾ documents:
      ▾ 0:
          doc_id:          "20683647"
        ▾ passages:
          ▾ 0:
            ▾ passage:     "We tested this hypothesis with larvae of the sequestering sphingid <species1>Ceratomia
                           catalpae</species1>, which is heavily <intbiot>parasitized by</intbiot> the endoparasitoid
                           <species2>Cotesia congregata</species2>, despite sequestering high concentrations of the iridoid glycoside
                           catalpol from their catalpa host plants."
              field:       "abstract"
  plazi:                   []
```

Figure 18: *Example of output for the biotic interaction triplets API.*

3.3. Evaluation

A benchmark has been created to assess the passage retrieval efficiency and passage retrieval effectiveness, following TREC guidelines (http://ciir.cs.umass.edu/million/guidelines.html , Pavlu and Aslam, 2007). This benchmark is based on the full biotic interactions generated for MEDLINE. Out of the passages returning a single result in MEDLINE, a set of 70 queries have been selected, through a manual screening in order to select relevant biotic interactions.

Focusing on queries returning only a single result can be regarded as simplification of the search tasks. It is, however, a well-established strategy - so-called know-item search - to develop a good search function when no gold standards are available, see (Voorhees, 2006).

3.3.1. Passage retrieval efficiency

The benchmark is used to query the biotic interaction triplets APIs and the resulting processing times have been recorded (Table 6). On average, it requires about 90 seconds, per query, to return results with the first version of the API, while the new API responds in 57 seconds. The demonstrator version of the GUI is based on the first version of the API.

Processing time (in seconds)	First version of API (s)	New version of API (s)
Total	6298.68	3971.88
Mean	89.98	56.74
Min	83.20	52.05
Max	98.16	60.99

Table 6: *Response times of the biotic interactions triplets API.*

Significant improvements are expected with the new version of the API, however we acknowledged that further developments will be needed to deliver friendly interactive services.

3.3.2. Passage retrieval effectiveness

The information retrieval tasks are performed in MEDLINE collection, via the SIBiLS. The proposed tasks are passage retrieval tasks, i.e. means that given a query, the engine is trying to recover a relevant passage. In the absence of an appropriate benchmark, comprising a query set, a document set and the relevance judgements linking the queries and the documents - see D11.1 - we decided to perform known-item search tasks. A known-item search is similar to an ad hoc search but the target of the search is a particular document (or a small set of documents) that the searcher knows to exist in the collection and wants to find again. Such a modelling is an approximation of real search tasks, which makes it possible to assess effectiveness of search engines when limited workforce is available to control for recall.

Benchmark

We generated 70 topics (annotations of biotic interactions), in t-uples forms: {species1,species2,interaction,pmid}. Species and interactions are normalised into a concept from OTT (OTT is the official labelling of the Open Tree of Life) or ROBI terminologies, and are provided with their preferred name (which is not necessarily the form present in the passage) along with a unique identifier. These 70 topics were controlled for consistency and manually selected to serve as "silver" standard to support the evaluation of the passage retrieval service.

An example of topics is given here :

```
{
  "topic_id": 3,
  "species_1_id": "OTT:857207",
  "species_1_name": "Yersinia pestis",
  "species_2_id": "OTT:844192",
  "species_2_name": "Bacteria",
  "biotic_interaction_id": "ROBI:0002626",
  "biotic_interaction_name": "kills".
  "pmid": "22593569"
}
```

Investigated strategies

Four strategies are investigated :

a) baseline : the names of the concepts (both species names and interaction) are used for an ad hoc search, i.e. a "Google-style" (e.g. Yersinia pestis Bacteria kills)
b) match_phrase: the names of the concepts are given within double quotes, and SHOULD be in the document (e.g. "Yersinia pestis" OR "Bacteria" OR "kills")
c) match phrase + AND: the names of the concepts are given within double quotes and MUST be in the article (e.g. "Yersinia pestis" AND "Bacteria" AND "kills")
d) annotations + AND: the IDs of the concepts MUST be in the annotations mapped by SIBiLS (e.g. "OTT:857207" AND "OTT:844192" AND "ROBI:0002626")

Metrics

Four different metrics are computed for evaluation, see (Manning 2008) for an introduction.

- Mean Reciprocal Rank (MRR) is the average inverse of the rank at which the relevant citation was returned.
- Recall at rank 10 (R10) is the percentage of topics for which the relevant document was returned in the top 10.
- Precision at rank 1 (P1) is the percentage of topics for which the relevant document was returned in rank 1.
- N is the average number of documents returned by SIBiLS, with N limited to 10000 by elasticsearch due to elasticsearch technical limitations.

Results and discussion

In this section, we report on the evaluations of the different explored retrieval tasks with the aforementioned metrics.

The discussion is focusing on MRR, which is mostly redundant with P1. MRR can be interpreted as the positional rank where the relevant match is found, so a MRR of 1 would mean that the relevant paper is returned in first position. Recall is here less interesting considering that the task is a known-item search task and not an ad hoc retrieval task.

	MRR	R10	P1	N
baseline	0.73	0.89	0.63	10000
match_phrase	0.73	0.87	0.64	5818
match_phrase + AND	0.77	0.83	0.71	1
annotations	0.95 (+23%)	1	0.9	1

Table 7: *Results for the passage retrieval task with complete queries. The number of MEDLINE records (N) returned by the passage retrieval service is provided in the last column.*

Using search engine's functionalities (such as match_phrase and AND operator) drastically reduce the number of documents retrieved (N) compared to the baseline. However, this does not lead to better performances in precision and recall metrics, probably due to the pre-existing ranking functions of the SIBiLS search engine. Fortunately, the annotations bring another level of performance: concepts not explicitly found in the text (because the author has used another form than preferred term) are retrieved thanks to the normalisation modules.

In complementary experiments, we computed the same metrics for incomplete queries : searching only with species (such as “Yersinia pestis” AND “Bacteria”) and searching with one species and the interaction (such as “Yersinia pestis” AND “kills”).

Complementary results : with only species
Here are results with only species used in query (not interaction) :

	MRR	R10	P1	N
baseline	0.63	0.86	0.51	4434
match_phrase	0.61	0.81	0.49	2144
AND + match_phrase	0.53	0.69	0.43	31
annotations	0.69 (+30%)	0.86	0.6	30

Table 8: *Results for the passage retrieval task with incomplete queries (only species).*

Complementary results : with one species and interaction
Here are results with only species 1 and interaction used in query (not species 2) :

	MRR	R10	P1	N
baseline	0.56	0.73	0.49	10000
match_phrase	0.57	0.71	0.5	5110
AND + match_phrase	0.73	0.87	0.63	6
annotations	0.78 (+6.5%)	0.93	0.67	6

Table 9: *Results for the passage retrieval task with incomplete queries (only one species and the interaction).*

As expected, the more specific the queries the higher the precision. The observation is however less valid for b) and c) than for d) strategies because there could be relevant passages but which are regarded as non relevant because the benchmark is made of full interactions. As expected, the normalisation (cf. line annotation) does improve the search effectiveness of the different search tasks. The gain is however limited when looking for species interacting with other species under a certain biotic modality (Table 9).

Further, we observe that searching for pairs of {species, interactions} is usually more effective than searching for pairs of species with respectively an MRR of 0.78 vs. 0.69. Such results suggest that searching for species interacting under a certain biotic relationship is likely to return more specific passages than searching for two species under any biotic interactions. It is yet to be confirmed how such differences will affect the final question answering tasks.

4. Conclusion and next steps

The passage retrieval service is ready, together with the prototype of the Graphical User Interface. Thanks to the combination of Plazi, MEDLINE and PMC contents, it has the potential to deliver the largest biodiversity knowledge library in JATS machine readable formats and additional journal corpora are likely to be added. The semantic normalisation layers, which are turning species synonyms into unambiguous named entities, are effective for passage retrieval but need intensive data cleaning. Further work is also needed to improve the response time of the biotic interaction explorer, up to the point where user interactions become fluent. The passage retrieval services and the biotic interaction explorer will serve as the basis for D11.4, the question answering services.

Furthermore, we need now to turn the developed services into production within the SIB Literature Services. In parallel, several integration scenarii with T11.2 (bidirectional linking services), which will be able to visualise interaction networks, are currently being considered.

Finally, we are also monitoring, through the progress of the GLOBI database and are coordinating with other biodiversity resources and projects (e.g. e-BioDiv, which is developing a service to bidirectionally curate specimen & citations cross-references) to leverage any relevant results for the BICIKL FAIR Data Place

5. Acknowledgements

We would like to thank the reviewers for their comments as well as the BICIKL consortium for the different feedback we received.

6. References

Gobeill, J., Caucheteur, D., Michel, P. A., Mottin, L., Pasche, E., & Ruch, P. (2020). SIB Literature Services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. *Nucleic Acids Research*, 48(W1), W12-W16.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Naderi, N., Mottaz, A., Teodoro, D., & Ruch, P. (2022). Analysing the Information Content of Text-Based Files in Supplementary Materials of Biomedical Literature. *Studies in Health Technology and Informatics*, 294, 876-877.

Pavlu, V., & Aslam, J. (2007). A practical sampling strategy for efficient retrieval evaluation. *College of Computer and Information Science, Northeastern University*.

Poelen, Jorrit, Upham, Nathan, Agosti, Donat, Ruschel, Tatiana, Guidoti, Marcus, Reeder, DeeAnn, Simmons, Nancy, Penev, Lyubomir, Dimitrova, Mariya, Csorba, Gabor, Groom, Quentin, & Willoughby, Anna. (2020). CETAF-DiSSCo/COVID19-TAF biodiversity-related knowledge hub working group: indexed biotic interactions and review summary (0.3) [Data set]. Zenodo.https://doi.org/10.5281/zenodo.4068958

Voorhees, E. M. (2007). Overview of TREC 2006. In *Proceedings of TREC 2006.* *https://trec.nist.gov/pubs/trec15/papers/OVERVIEW.pdf*