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

Deliverable D11.1

29 April 2022

Déborah Caucheteur[16], Alexandre Flament[16], Sofie Meeus[4], Wouter Addink[2,3], Christos Arvanitidis[5], Bachir Balech[6], Mathias Dillen[4], Mariya Dimitrova[7,8], Juan Miguel González-Aranda[5], Jörg Holetschek[9], Sharif Islam[2,3], Thomas S. Jeppesen[10], Daniel Mietchen[11,12,13], Nicky Nicolson[14], Lyubomir Penev[8], Tim Robertson[15], Maarten Trekels[4], Olaf Bánki[2], Quentin Groom[4], Donat Agosti[1], Emilie Pasche[16], Patrick Ruch[16]

Authors' affiliations:

[1] *Plazi, Bern, Switzerland*

[2] *Naturalis Biodiversity Center, Leiden, Netherlands*

[3] *Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands*

[4] *Meise Botanic Garden, Meise, Belgium*

[5] *LifeWatch ERIC, Seville, Spain*

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.

[6] *Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, CNR, Bari, 70126, Italy*

[7] *Bulgarian Academy of Sciences, Sofia, Bulgaria*

[8] *Pensoft Publishers, Sofia, Bulgaria*

[9] *Botanic Garden & Botanical Museum Berlin-Dahlem, Berlin, Germany*

[10] *Danish Natural History Museum, Copenhagen, Denmark*

[11] *EvoMRI Communications, Jena, Germany*

[12] *University of Virginia, Charlottesville, United States of America*

[13] *Data Science Institute, University of Virginia, Charlottesville, United States of America*

[14] *Biodiversity Informatics & Spatial Analysis, Royal Botanic Gardens, Kew, London, UK*

[15] *Global Biodiversity Information Facility, Copenhagen, Denmark*

[16] *HES-SO & SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	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.
Deliverable n°:	D11.1
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP11
Lead beneficiary:	Plazi
Citation:	Caucheteur, D., Flament, A., Meeus, S., Addink, W., Arvanitidis, C., Balech, B., Dillen, M., Dimitrova, M., González-Aranda, J., Holetschek, J., Islam, S., Jeppesen, T., Mietchen, D., Nicolson, N., Penev, L., Robertson, T., Trekels, M., Bánki, O., Groom, Q., Agosti, D., Pasche, E., Ruch, P. (2022) *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*. Deliverable D11.1 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 12
Actual submission date:	29 April 2022

Deliverable status:

Version	**Status**	**Date**	**Author(s)**
1.0	Draft	March 2022	WP11 members
2.0	Review	April 2022	Q. Groom, B. Barov
3.0	Submission	2022	-

Table of contents

Executive summary 5
Introduction 6
Methodology 7
Benchmark 9
Factoid questions (e.g. Wh-questions such as where, when, who, ...): questions which can be answered with a concept of a short phrase 9
Biotic interaction 9
Fact extraction 9
Taxonomic 9
Specimen and taxonomic relationship / Museum or Botanic gardens (~biobanks) 10
Geographic location 10
Occurrences 10
Literature citation 10
Cell lines 10
Genomics 10
Drugs 10
Open questions: causal question (why ?, how ?) 11
Biotic interaction 11
Etiology 11
Resource management, biodiversity conservation 11
Climate-change impact, responses to climate change 11
Results 12
a broad-coverage list of biodiversity resources 12
A prioritised list of factoid questions 13
Relevance judgements 14
Infrastructures 14
Recommendations 15
Qualitative analysis of questions 15
Document corpus 16
Implementation workflows 17
Case #1: Mutation search in supplementary data 17
Case #2: Complex biotic interactions 20
Conclusion 25
Acknowledgements 25
References 25
Appendix 27
Appendix A 27
Factoid questions (e.g. Wh-questions such as where, when, who, ...): questions which can be answered with a concept of a short phrase 27
Open questions: causal question (why ?, how ?) 36
Appendix B 39

Executive summary

Background: The BiCIKL project is born from a 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 infrastructures, particularly for 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. The FAIR Data Place is thus described as a powerful information retrieval platform likely to answer a wide range of biodiversity-related questions Because designing an "universal" search system through various research infrastructures holding different data types is not realistic, D11.1 aims at collecting a set of user questions and answers to those questions. From this broad set a subset will be selected to form the basis of future services to be implemented (T11.2-5). The main result of this effort is a data asset consisting of a set of user information needs as expressed via natural language questions, each supplied with a set of evidence-based answers.

Methods: Leveraging metrology methods developed by the US National Institute of Technology (NIST), we report on the content of the delivered benchmarks (i.e. questions, answers and links to evidence) and briefly describe the evaluation methodology, including metrics, used in the field.

Results: A total of N=31 questions and answers have been gathered. A curation effort has been started to classify and prioritise these questions & answers according to categories such as: factoid questions vs. open questions, complex vs. single database answers, semantic types (e.g. biotic interactions, biospecimen, geographic locations). The quantitative analysis shows the heterogeneity of the databases as well as the prevalence of queries related to biotic interactions. Further, two question examples - one complex and one simple - have been selected to derive different implementation workflows in order to support subsequent developments of WP11.

Conclusion: The benchmark is ready to support the prototyping and evaluation of future WP11 developments. The question survey suggests that biotic interactions could play a federative role to design future WP11 services.

1. Introduction

The overarching goal of BiCIKL is to create a community of infrastructures concerned with data on biodiversity through liberating data from scholarly publications and bi-directional linking of literature, taxonomic, DNA sequence and occurrence data (Penev et al. 2022). By working together, linking data, practising Open Science and Open Innovation, the project aims to make biodiversity data much more accessible and particularly to make these data more interoperable with the ultimate vision of making them more useful for novel research and informed policy decisions. In addition to the Open Science aspect of BiCIKL, there are also the good practices for data management that are summarised in the FAIR Data Principles (Wilkinson et al. 2016). These principles are a guide to how to make data more *findable*, *accessible*, *interoperable* and *reusable*. Open Data is not a prerequisite for complying with the principles, but does often make compliance considerably easier. Certainly, the FAIR Data principles include having the metadata - describing the data - open as a prerequisite for findability.

At a technical level BiCIKL intends to achieve its goals through the provision of data, tools and services to the community. It will cover the whole research life cycle and will contribute new methods and workflows to harvest, liberate, link, reuse data from specimens, samples, sequences, taxonomic names and taxonomic literature (Figure 1). Yet, both the technology and the community need to align with this vision, and hackathons can be a means to ensure this alignment.

WP11 in particular aims at designing and developing a “FAIR Data Place”, i.e. an improved integration service and portal to support the interlinking and distribution of queries across Biodiversity databases (Figure 1). Considering that providing an universal query distribution and interlinking mechanism likely to interface and connect all biodiversity data sources is not a realistic endeavour, we instead hope to prioritise a few high importance use cases, exemplified with concrete user information requests. While the definition of use cases is explored by different deliverables (e.g. Hackathon, Questionnaire), D11.1 focuses on the delivery of a question answering benchmarks. A set of questions have therefore been collected from the participants of the project. Each question was to be provided with additional information such as one or several answers - including some accession numbers (or catalogue number) when found in a database - and a link to evidence supporting the answer(s) - including accession numbers when available.

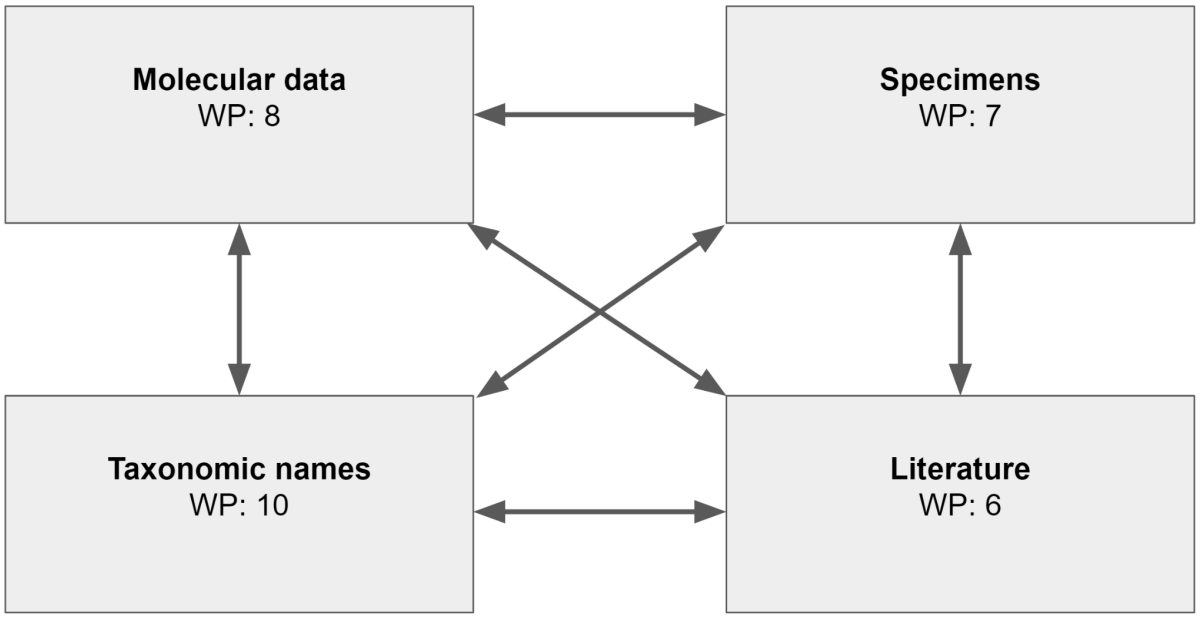

Fig.1: *Diagram of entities and relationships across biodiversity data endpoints. A total of six (pair-wise) relationships can be obtained.*

2. Methodology

The deliverable D11.1 is strictly following the so-called Cranfield model, conceptualised by the US National Institute of Standards of Technology[1] via the Text Retrieval Conferences (TREC). Originally designed in the 80's and 90's to evaluate emerging textual search engine technologies, the model has been gradually expanded to cover any search tasks, including multimodal search in virtually any contents, including web, social media, images or video.

TREC-like benchmarks are made of three major components (see triangle in Figure 2):

1. a set of queries,
2. a set of documents,
3. a set or relevance judgements.

The set of queries (or questions for question answering engines) is ideally in the range of N=25-50, while the set of documents is any large corpus. Typical corpora contain more than a million documents or data elements. The relevance judgements are curated links which connect a given query with a given document (or answer for question answering). Several search tasks have been explored during the lifespan of TREC (~50 years) - see section 3.3 – but the methodology and metrics (section 3.3) remain extremely stable and robust.

[1] https://www.nist.gov/; https://trec.nist.gov/

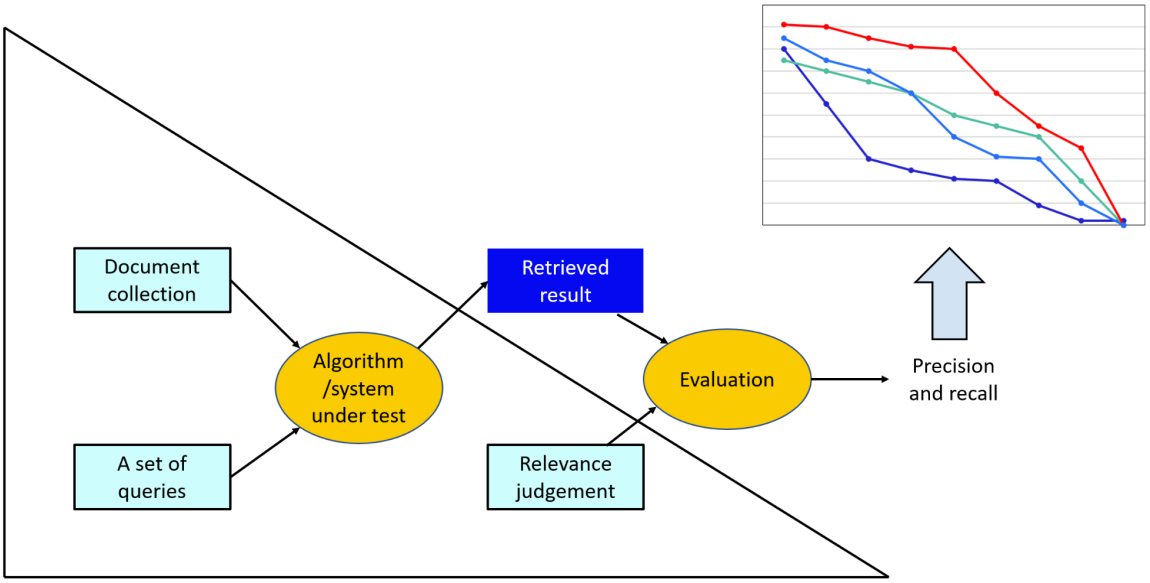

Figure 2: *Methodological workflow of information retrieval experiments to support the design and evaluation of WP11's data discovery services*[2].

In this report, we describe the results of the benchmark development process, from which we derive several architectural use cases to support the development of the FAIR Data Place, i.e. we focus on the triangle from Figure 2. The evaluation will be the subject of future tasks of WP11.

First, we describe the generated benchmark. Second, we analyse the contents of the benchmark with reference to the main databases - including the document corpus - needed to implement the question answering tasks. Third, we introduce the most common information retrieval measures (section 3.3). Fourth, we prioritise a few questions, from which we derive different answering scenarii. Further, a set of conceptual diagrams are delivered to showcase the implementation work.

2.1 Sample of factoid questions

The next paragraph shows a sample of the questions. The complete list is available in the Appendix A at the end of the report.

[2] After D. Glowacka https://glowacka.org/lectures/ir/interactive-IR2.pdf

Benchmark

Factoid questions (e.g. Wh-questions such as where, when, who, ...): questions which can be answered with a concept of a short phrase

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?

Fact extraction

1.20 What viruses are shared between bat x and y?
1.21 What is the geographic distribution of the bats that share the highest number (same species) of viruses?
1.23 What biotic interaction is used by *Dermatobia hominis* to trigger myiasis in humans?

Taxonomic

1.8 What is the currently accepted name for all derivative components of specimen collection event Seigler 16161 on 27 May 2007?
For example (but not this specific collection event), a botanist collects three duplicates from a tree. These get sent to three collections and get differently curated. Data related to this collection event is now in three collections, ENA and cited in the literature. A taxonomic expert reidentified a specimen (CoL) at one of the herbaria. What name is the one that was placed on the data by taxonomic experts? How can that update be sent to the other components?
1.11 What are the parents of the hybrid "Solanum × michoacanum"?
1.16 What is the phylogenetically closest living relative of the Dodo (*Raphus cucullatus*)?

Specimen and taxonomic relationship / Museum or Botanic gardens (~biobanks)

1.10 Which collection can provide access to specimens of the genus *Latrodectus*?
1.12 Where can I find a specimen of Dodo *Raphus cucullatus*?

Geographic location

1.9 At which altitude was found the holotype of Phragmataecia Newman, 1850?
1.2 What is the origin of SARS-Cov-2?

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 ?

Literature citation

1.13 Who discovered the species Dodo *Raphus cucullatus*?
1.14 Which articles describe species Dodo *Raphus cucullatus*?

Cell lines

1.7 What cell lines are used to study SARS-CoV-2?

Genomics

1.4 What are human genes involved in Covid-19 infections?
1.5 What are the main Variants of Concern for SARS-Cov-2?

Drugs

1.6 What drugs have been active against SARS-CoV-2 in animal studies?

Open questions: causal question (why ?, how ?)

Biotic interaction

2.1 How do raccoons impact the population size of bats in Europe?
2.5 Which spiders do *Sceliphron* wasps predate?
2.7 What insects are hosted by a particular plant? Example of a use case for this information will help species conservation by guiding the public to improve biodiversity in their gardens and guiding planting for conservation purposes.

Etiology

2.2 What is the origin of SARS-Cov-2 ?

Resource management, biodiversity conservation

2.3 As a Conservation Planner, I want to cross-check species identification against reliably identified specimens to create a checklist of species. For this, data such as images, sequence data, georeferences and traits are needed.

Climate-change impact, responses to climate change

2.4 As a researcher, scientist I want to search for trait information of a certain species and answer the question: How do species traits change based on changes in the environment due to global warming? Digitized collections, location, date, high-resolution images of specimens, and trait information are needed.
2.6 How does the coronavirus respond to changes in the weather ?

2.2 Quality control of the collected benchmark

From this list of questions, we can also identify a few very generic templates (e.g. What viruses are shared between bat x and y?), which will need further work to become proper information requests associated with some non-ambiguous answers.

Further, we observe that some of the collected questions seem ambiguous: thus, question 2.2. is relatively unclear as it could refer to different aspects such: 1) as the geographic origin or 2) the chain of causality, which triggered the pandemics (e.g. a chain of biotic interactions). These aspects are only a subset of the many different interpretations possible for these questions. However, in many situations, the ambiguity is solved as soon as we consider the answer supplied together with the question by the authors. Here the authors expect the following answer: “Wu-Han, China”, i.e. the geographic location, which clearly discards the many alternative interpretations. Similarly some of the self categorised questions (e.g. 2.5) were likely not assigned to the right question category (e.g. open vs. factoid questions).

It is worth observing that the methodology as designed by the NIST is an effort to provide a realistic set of questions and answers; it is therefore expected that some of the questions may have different interpretations or may even be sub-optimally formulated. Future steps could lead to rejecting some of the questions but the initial collection step aims at collecting a raw material, as shown in the list of questions and answers.

3. Results

In this section we report on the main observation derived from the Question Answering benchmarks. We also describe the corpora statistics.

3.1. a broad-coverage list of biodiversity resources

The following provides the list of resources covered by the answer field of the benchmark.

- Biodiversity Literature Repository (BLR)
  https://biolitrepo.org/
- Catalogue of Life (COL)
  https://www.catalogueoflife.org/
- European Nucleotide Archive (ENA)
  https://www.ebi.ac.uk/ena/browser
- Encyclopedia of Life (EOL)
  https://eol.org/
- Global Biodiversity Information Facility (GBIF)
  https://www.gbif.org/
- Plazi

https://plazi.org/ *
- TreatmentBank
  https://plazi.org/treatmentbank/
- SIBiLS
  https://candy.hesge.ch/SIBiLS/
- The Open Biodiversity Knowledge Management System (OpenBioDiv)
  https://openbiodiv.net/
- Meise Botanical Garden (MBG)
  https://www.botanicalcollections.be
- Botanic Garden and Botanical Museum (BGBM)
  https://www.bgbm.org/en/biodiversity-informatics
- PlutoF
  https://plutof.ut.ee/
- Lifewatch - Catalogue of Virtual Labs
  https://www.lifewatch.eu/catalogue-of-virtual-labs/
- Cross-Ref
  https://www.crossref.org/
- EuropePMC
  https://europepmc.org/
- Global Biotic Interactions (GLOBI)
  https://www.globalbioticinteractions.org/
- Medical Literature Analysis and Retrieval System Online (MEDLINE)
  https://www.nlm.nih.gov/medline/index.html
- Morphbank
  https://www.morphbank.net/
- NCBI Blast GenBank
  https://blast.ncbi.nlm.nih.gov/Blast.cgi
- NCBI taxonomy
  https://www.ncbi.nlm.nih.gov/taxonomy
- neXtProt
  https://www.nextprot.org/
- SynoSpecies
  https://synospecies.plazi.org/
- UniProt
  https://www.uniprot.org/

3.2. A prioritised list of factoid questions

The full content of the question collecting process is available in appendix A.

This documentation was provided to the respondents to support the question acquisition process, which has been running since the beginning of the project.

The main result of the collection process is a list of 31 questions and answers, documented with the evidence supporting the answer. The volume is consistent with TREC (Text Retrieval Conferences) recommendations.

3.3. Relevance judgements

Out of appendix A, a set of relevance judgements will be generated, complying with TRECeval formatting standards. The relevance judgements will allow to cover different evaluation tasks such as ad hoc retrieval, known-item search and ultimately question-answering. Recall and precision metrics - as well as weighted means of these two dimensions - will be defined based on the different tasks as shown in table 1.

num_q	Total number of evaluated queries
num_ret	Total number of retrieved documents
num_rel	Total number of relevant documents (according to the qrels file)
num_rel_ret	Total number of relevant documents retrieved (in the results file)
map	Mean average precision (map)
gm_map	Average precision. Geometric mean
Rprec	Precision of the first R documents, where R are the number os relevants
bpref	Binary preference
recip_rank	Reciprical Rank
iprec_at_recall_0.00	Interpolated Recall - Precision Averages at 0.00 recall
iprec_at_recall_0.10	Interpolated Recall - Precision Averages at 0.10 recall
iprec_at_recall_0.20	Interpolated Recall - Precision Averages at 0.20 recall
iprec_at_recall_0.30	Interpolated Recall - Precision Averages at 0.30 recall
iprec_at_recall_0.40	Interpolated Recall - Precision Averages at 0.40 recall
iprec_at_recall_0.50	Interpolated Recall - Precision Averages at 0.50 recall
iprec_at_recall_0.60	Interpolated Recall - Precision Averages at 0.60 recall
iprec_at_recall_0.70	Interpolated Recall - Precision Averages at 0.70 recall
iprec_at_recall_0.80	Interpolated Recall - Precision Averages at 0.80 recall
iprec_at_recall_0.90	Interpolated Recall - Precision Averages at 0.90 recall
iprec_at_recall_1.00	Interpolated Recall - Precision Averages at 1.00 recall
P_5	Precision of the 5 first documents
P_10	Precision of the 10 first documents
P_15	Precision of the 15 first documents
P_20	Precision of the 20 first documents

Table 1: *Metrics to evaluate search and questions-answering tasks of WP11.*

3.4. Infrastructures

Data science often requires large amounts of data to be analysed, and the only way to process this data efficiently is to create a local copy. Infrastructures (Figure 3) should provide download access to all or part of the data so that it can be processed remotely by researchers. This could be provided in several ways. GBIF provides an asynchronous download system for queries and direct downloads of individual datasets. In the absence of a

dedicated download system, users may try to achieve the same result through an API, but this is highly inefficient for the user and infrastructure.

Recommendations

- Provide as many different modalities of access as possible.
- Avoid requiring personal contacts to download data.
- Provide a full description of an API and the data it serves

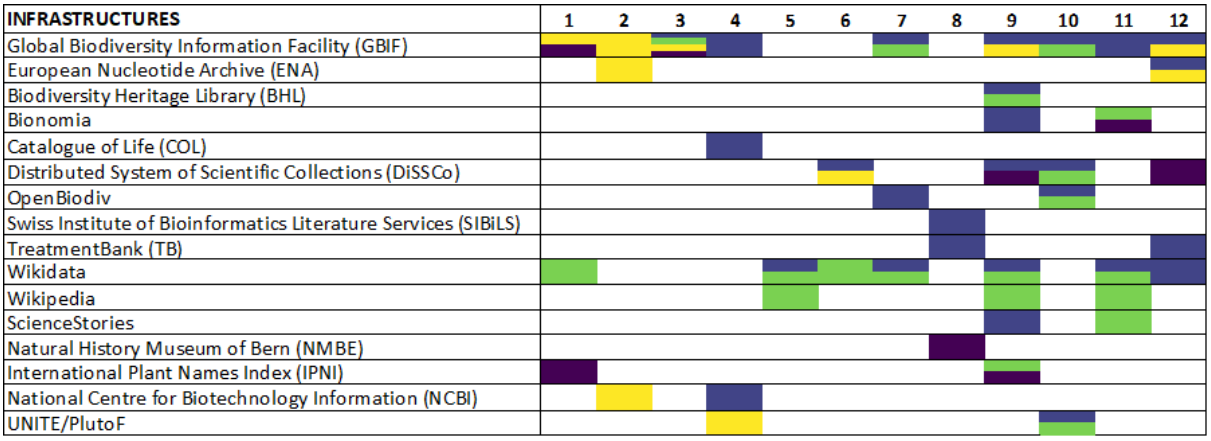

Fig. 3: *The modes of access to the different infrastructures used by hackathon project teams: = Application Programming Interface or API (eg. SPARQL, RestFul); = website, manual access; = download or dump; and = personal request. The numbers of columns refer to the hackathon sub-projects (reference, link, or citation in press)*

3.5. Qualitative analysis of questions

From the analysis of the questions set, we can roughly identify two question types: 1. (simple) questions likely to be answered from one knowledge source, 2. (complex) questions involving at least two knowledge sources. A third category would involve questions in RDF powered with a SPARQL endpoint. However, in the datasets only a unique source of data was provided with such a powerful query algebra. With these three sets, we likely cover all questions from the biodiversity community. While simple questions are likely to be answered by each database provider, and SPARQL endpoints seem rare in the field, the next steps will be to prioritise a subset of questions (e.g. 5) with high impact for the community to demonstrate the FAIR Data Place with a special focus on biotic interactions questions, which account for more than half of the questions in the benchmark.

3.6. Document corpus

In addition to the databases listed in section 3.1, table 3 shows statistics of the corpora available to answer the questions: primarily MEDLINE and PMC, including authors' manuscripts. Specific sub-collections are also currently being harvested (e.g. Plazi). Similarly, a growing number of annotations are being generated with currently more than 2 billion available within SIBiLS, the SIB Literature Services. Of particular interest for the benchmarks are annotations related to biotic interactions, which account for more than 50% of the complex factoid questions of the benchmarks, see example in Figure 4.

Type	**Terminology used**	**Nb of entities annotated in MEDLINE**	**Nb of entities annotated in PMC**
Drugs	DrugBank	79 202 305	76 537 408
Drugs	ATC	45 695 077	26 488 868
Disease	NCIt	129 587 276	91 884 424
Disease	ICD-O3	4 767 683	2 467 524
Gene	NeXtProt/Uniprot	35 649 221	83 204 241
Functions	Gene Ontology	42 996 047	77 322 206
Medical entities	MeSH	415 097 413	612 833 130
Evidences	ECO	4 269 843	11 042 888
Species	NCBI Taxonomy Browser	9 914 247	16 718 026
Nb total documents:		*30 008 991*	*2 374 281*
Nb total annotations:		***773 427 866***	***1 009 850 920***
Average per document:		*26*	*425*

Table 2: *More than 30 million MEDLINE articles and 4.5 million full-text articles have been gathered to complete the evaluation corpora. The collection will grow thanks to the planned harvesting of Plazi's treatments.*

Bulinus	host of	Schistosoma haematobium	[304396]. Ecological studies of Bulinus rohlfsi, the intermediate host of Schistosoma haematobium in the Volta Lake.
Panstrongylus	host of	Trypanosoma cruzi Panama	[328884]. Geographical extension in a new ecological association of Panstrongylus humeralis (Hemiptera: Reduviidae) natural host of Trypanosoma cruzi in Panama.
Ooencyrtus	parasite of	Rhodnius prolixus Vector	[331437]. [Effect of parasite density of Ooencyrtus trinidadensis (Chalcidoidea, Encyrtidae), an endophagus parasite of eggs of Rhodnius prolixus, vector of Chagas' disease in Venezuela].

Fig. 4: *Example of NCBI Taxonomy annotation of MEDLINE and PMC with biotic interactions based on the OBO relation ontology.*

4. Implementation workflows

In the following, we show how some of the simple or complex questions could be answered via the FAIR Data Place. We picked two questions as follows:

1. Case #1: simple question, likely to be answered via one knowledge source
2. Case #2: complex question, which needs more than one knowledge source to be answered and which can be answered via different channels.

These two examples show how WP11 services could be modelled and implemented. It is worth noting that case #2 leverages three biodiversity entities out of the four entities shown in Figure 1; namely: sequences, literature, and taxonomic names.

4.1. Case #1: Mutation search in supplementary data

Give me all papers where human ABL1(F359V) is mentioned in supplementary data images

This question fits exactly the Variomes service (https://candy.hesge.ch/Variomes). Variomes is a high recall search engine supporting the curation of genetic variants. It enables users to search variants in various collections such as full text articles (PubMed Central), and the supplementary data associated with these articles. This system uses a variant synonym generator to increase the comprehensiveness of retrieved documents. The query plan requires a unique call:

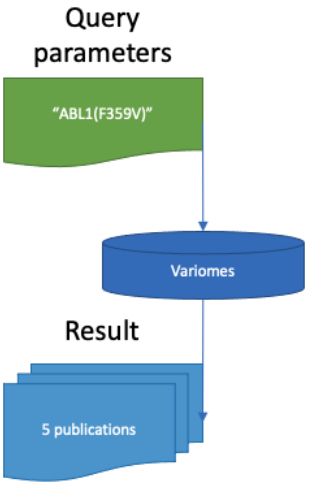

Fig. 5: *Query plan for the case #1.*

Step 1: enter gene or gene product with variant (at protein, transcript or DNA level) in the landing page of Variomes.

Fig. 6: *Landing page of Variomes.*

Step 2: display of the retrieved files

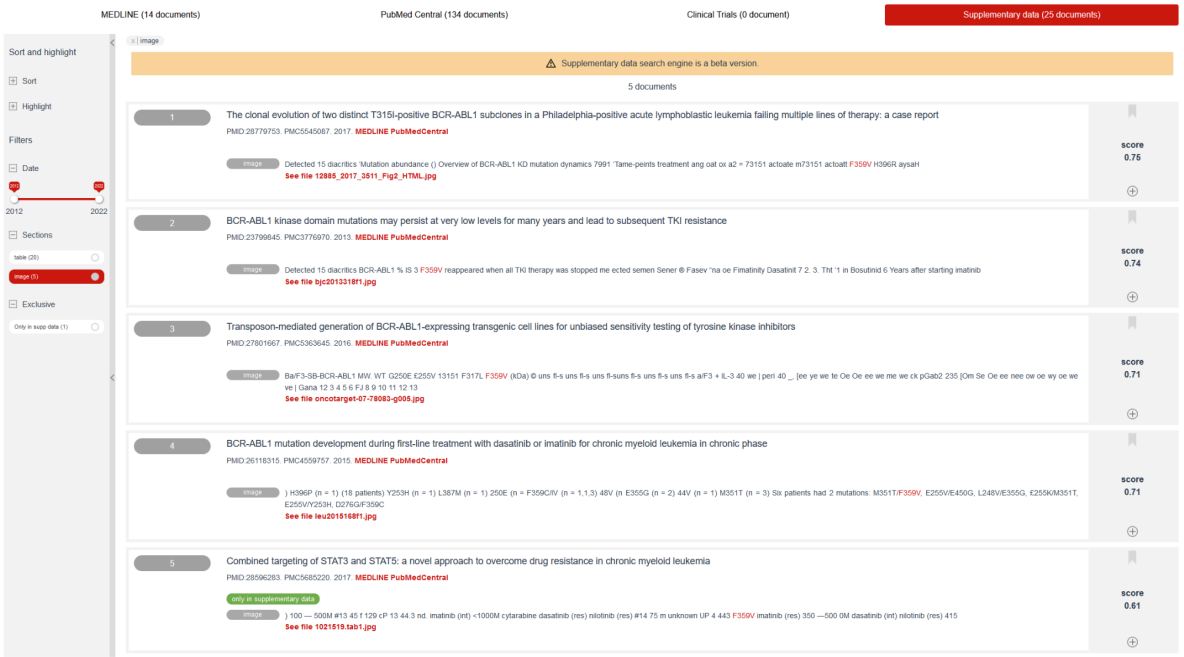

Fig. 7: *Results returned by Variomes with a tab on the left showing contents in Supplementary Data files and facets on the right selected to show up only images.*

Some examples of the fetched images are found below
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5545087/

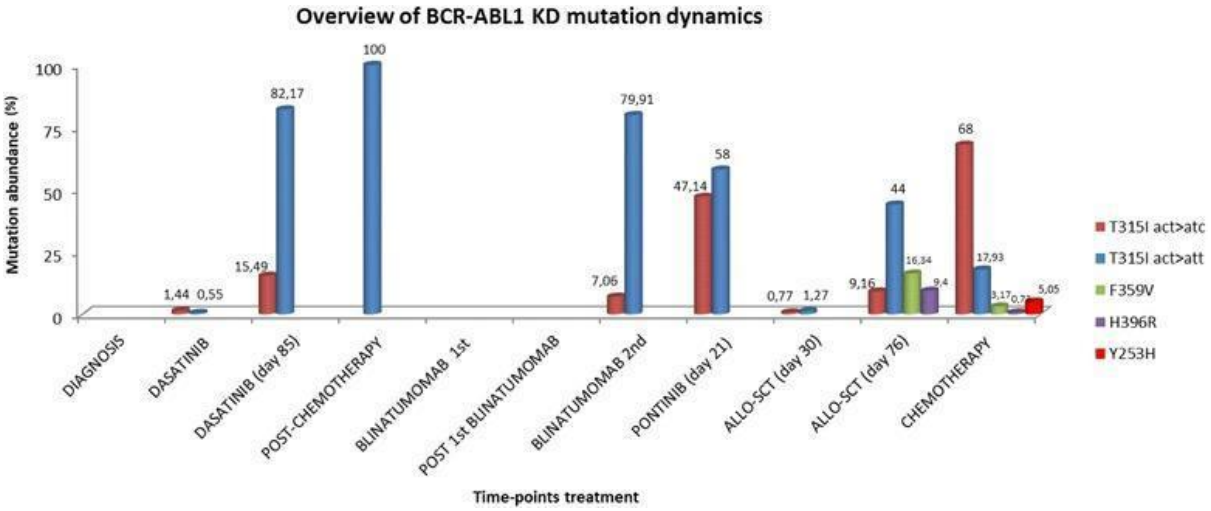

Image 1. *First result from Variomes*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3776970/

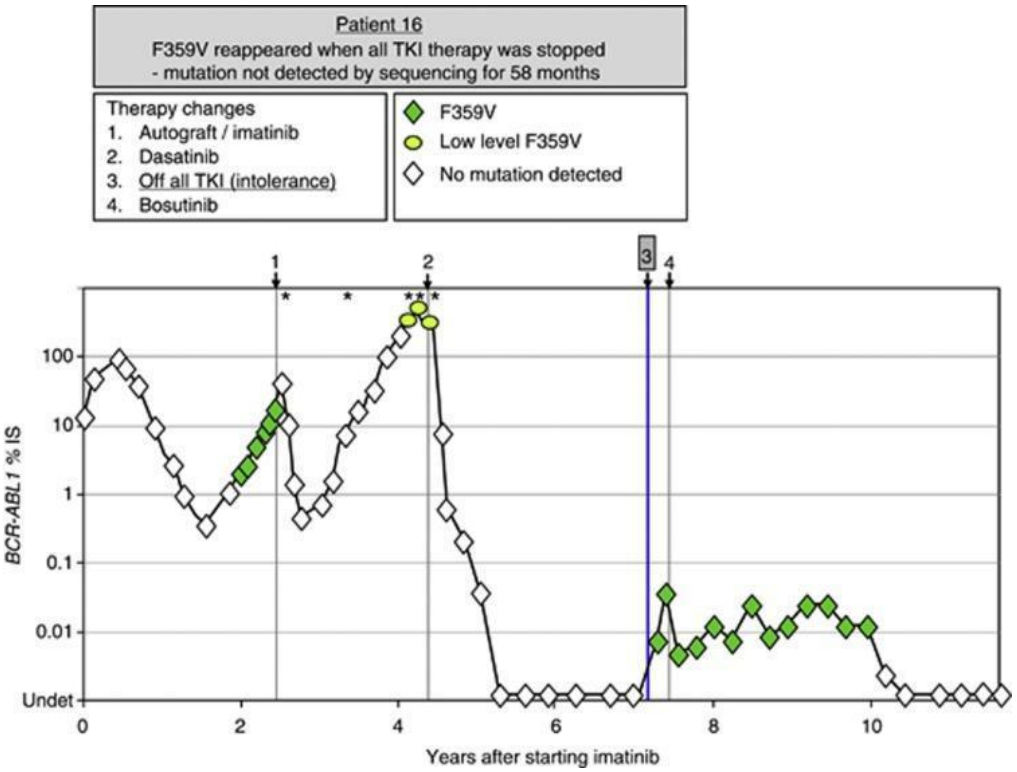

Image 2. *Second result from Variomes: F359V is found at the top of the image.*

4.2. Case #2: Complex biotic interactions

Find which species interact with the sequence owner (below) and is involved in Myiasis

```
ATGAATAAACCTTTACGAATTAAACACCCAATTTTCAAAATTGCTAATAATGCACTAATT
GATCTACCAGCTCCTATTAATATTTCTGCATGATGAAATTTTGGATCTCTTCTTTTTTTA
TGTTTAATAATCCAAATCTTAACTGGACTATTTCTAGCCATACATTACACAGCAGATATT
AATTTAGCATTTAATAGAGTTAATCATATCTGCCGAGATGTAAATTATGGATGATTATTA
CGAACAATACATGCCAACGGTGCATCATTCTTTTTCATTTGTATTTATTTACATGTAGGA
CGTGGAATTTATTATGGATCTTACCTTTTTTCACCAACATGATTAATTGGTGTAATTATC
CTATTTTTAGTAATAGGTACAGCTTTTATAGGTTATGTATTACCATGAGGACAAATATCC
TTTTGAGGAGCTACAGTAATTACAAATTTATTATCAGCCATCCCATATTTAGGAATTGAT
TTAGTACAATGAGTATGAGGAGGATTCGCCGTAGACAATGCAACATTAACTCGATTTTTT
ACTTTTCATTTTATCCTCCCATTTATTGTACTAGCTATAACTATAATTCATATTTTATTT
TTACATGAAACAGGATCCAATAATCCTATAGGATTAAATTCAAATACTGATAAAATTCCA
TTTCATCCATATTTTACTTTTAAGGATATCGTAGGATTTATCGTAATAACAGCAATCTTA
ATTATATTAGTTTTAATTAATCCATATCTATTGGGAGACCCAGATAATTTTATTCCAGCT
AATCCATTAGTTACCCCCGTTCACATTCAACCAGAATGATATTTTTTATTTGCTTATGCT
ATTCTTCGATCAATTCCTAATAAATTAGGAGGAGTAATTGCTCTAATTCTATCAATTGCT
ATTTTAGCAATTCTCCCATTCTATAATTTAAGTAAATTTCGAGGAATTCAATTCTACCCA
ATTAATAAATTAATATTTTGAATAATAACTATTACAGTAATTTTATTAACATGAATTGGA
GCTCGACCTGTAGAAGAACCATATGTACTAGTGGGACAAATTCTAACAGTATTATATTTT
  TCTTATTTCTTATTAAACCCAATAATTACAAAATGATGAGATAATTTACTAAATTAG
```

For such complex questions, there is no service providing direct answers. We thus propose here a pipeline (Figure 8) consisting of multiple chained queries:

1. Identification of the species owning the given sequence.
2. Identification of a list of species interacting with the species identified in step 1.
3. Filtering the list of species retrieved in step 2 to select only the species involved in myiasis.

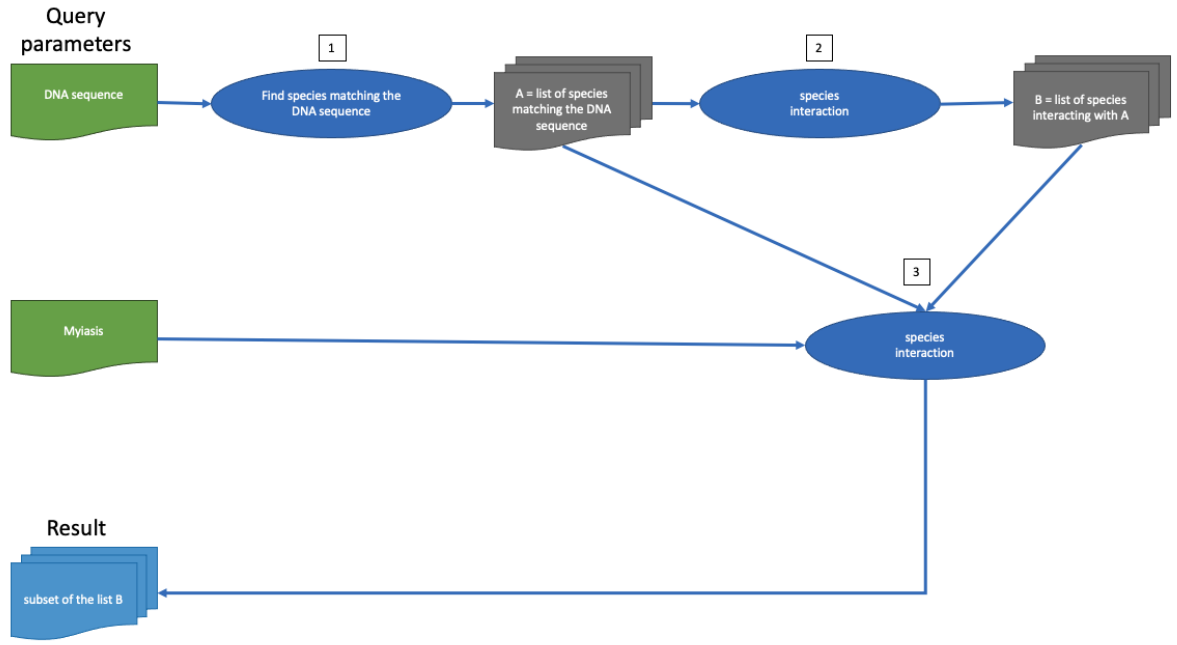

Fig. 8: *Multiple chained queries pipeline to answer complex query for case #2.*

In the following, we are describing in more details the three steps. A general pipeline including the services is proposed in figure 9.

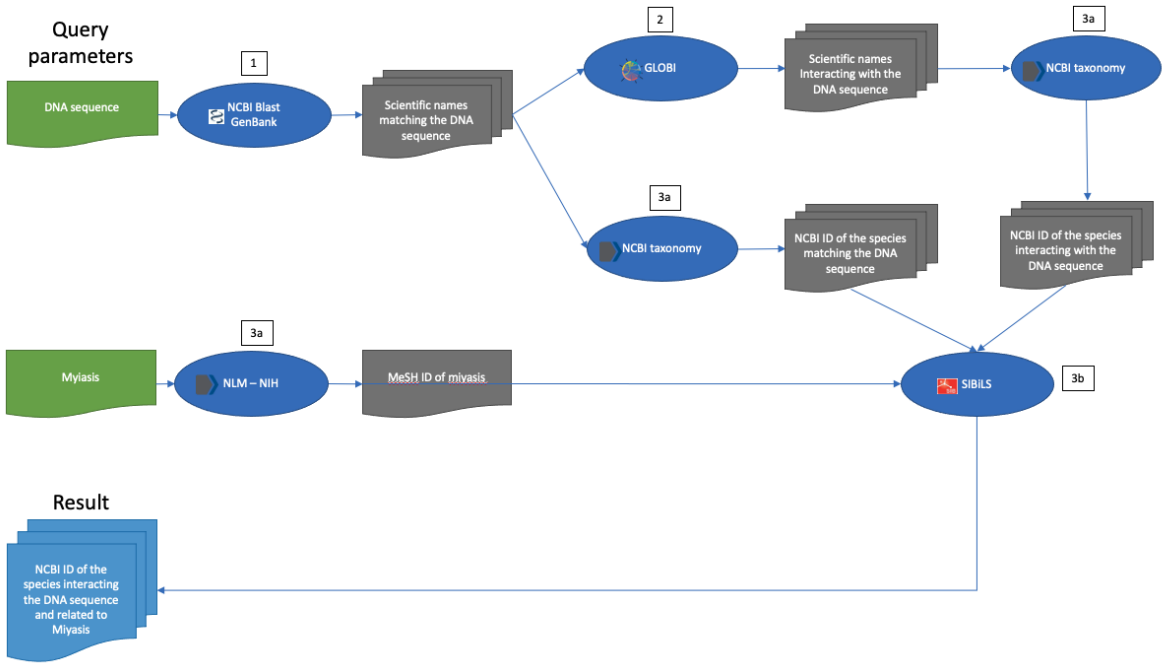

Fig. 9: *Multiple service call pipeline to answer complex query for case #2.*

Step 1.

The first step consists of identifying which species include in their genes the DNA sequence. To achieve this, a primary local alignment search tool is used (i.e. BLAST). Such tools compare

nucleotides of a given sequence to sequences in a database and return potential candidates based on the statistical significance. We propose to use the GenBank database to search for similar sequences. The GenBank BLAST service is publicly available at https://blast.ncbi.nlm.nih.gov/Blast.cgi .

As an output of this service, a list of sequences producing significant alignments is suggested. It can thus produce several candidates to a given sequence. In our example, *Dermatobia hominis* results in a 100% alignment with our sequence (Figure 10).

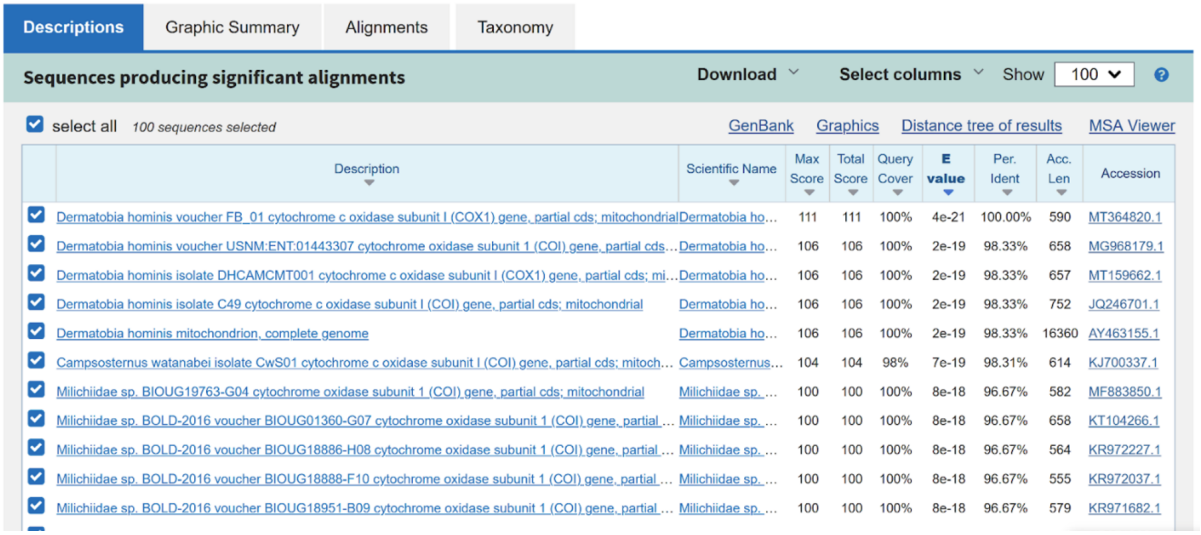

Fig. 10: *Results from the GenBank BLAST service.*

Step 2

The second step consists of identifying a list of species interacting with the species identified in step 1 (i.e. *Dermatobia hominis*). It is to be noted that step 2 must be repeated for each species identified in step 1. We propose to use GLOBI (https://www.globalbioticinteractions.org), a service providing interaction between species based on the combination of open datasets.

As an output of this step, a list of 8 distinct species is obtained (Figure 11).

download csv data sample		access full dataset
taxon	interacts with	taxon
(2 distinct)	(10 distinct interactions)	(8 distinct)
Dermatobia hominis	parasite of	Blastocerus dichotomus
Dermatobia hominis	parasite of	Bos taurus
Dermatobia hominis	parasite of	Homo sapiens
Dermatobia hominis	parasite of	Saguinus mystax
Dermatobia hominis	parasite of	Alouatta palliata
Dermatobia hominis	interacts with	Psorophora ferox
Dermatobia hominis	interacts with	Proteus mirabilis
Dermatobia hominis	parasite of	Homo sapiens
Dermatobia hominis	has host	Homo sapiens
Dermatobia hominis	parasite of	Canis familiaris

Fig. 11: *Result from the GLOBI service*

Step 3.

The final step consists of identifying which of the 8 species are involved in myiasis. To this extent, we propose to search the literature for co-occurrences of each of these species with *Dermatobia hominis* and myiasis. We propose to use the SIB Literature Services (SIBiLS). SIBiLS is enriched with a set of nearly 2 billion of mapped biomedical entities from reference vocabularies. Using this annotation of the content enables to improve the recall of the queries: indeed, it not only retrieves the exact search term, but also its synonyms and syntactic variations.

Thus, as a preliminary step (3a) to querying SIBiLS, we first need to map the species (i.e. *Dermatobia hominis* and the 8 species identified in step 2), as well as the target disease (i.e. myiasis) to reference vocabularies. We suggest using NCBI taxonomy for species and MeSH for diseases. MeSH is already present in the SIBiLS reference vocabularies, while NCBI taxonomy is currently under process and will be soon available. The mapping can be done automatically using NCBI taxonomy and MeSH APIs. For instance, *Marsh deer* will be mapped to NCBI taxon 248133.

When all species are mapped, we can search for co-occurrences in SIBiLS (3b). The SIBiLS APIs are used: candy.hesge.ch/SIBiLS/MEDLINE/search.jsp for MEDLINE abstracts and candy.hesge.ch/SIBiLS/PMC/search.jsp for full-text articles. We query these APIs with a set of triplet (i.e. a triplet consists of one of the species retrieved in step 2, the species identified in step 1 and the disease mentioned in the query). If no document is retrieved for a triplet, we

assume that the species identified at step 2 is not involved in myiasis with *Dermatobia hominis*.

As shown in the figure 12, no publication has been found matching *Myiasis*, *Dermatobia hominis* and *Marsh deer*; all concepts unambiguously identified via either NCBI Taxonomy Accession Numbers or Medical Subject Headings.

Fig. 12: *Result from the SIBiLS service for "Marsh deer".*

However, thirty papers contain the triplet *Myiasis*, *Dermatobia hominis* and *Bos taurus* (figure 13).

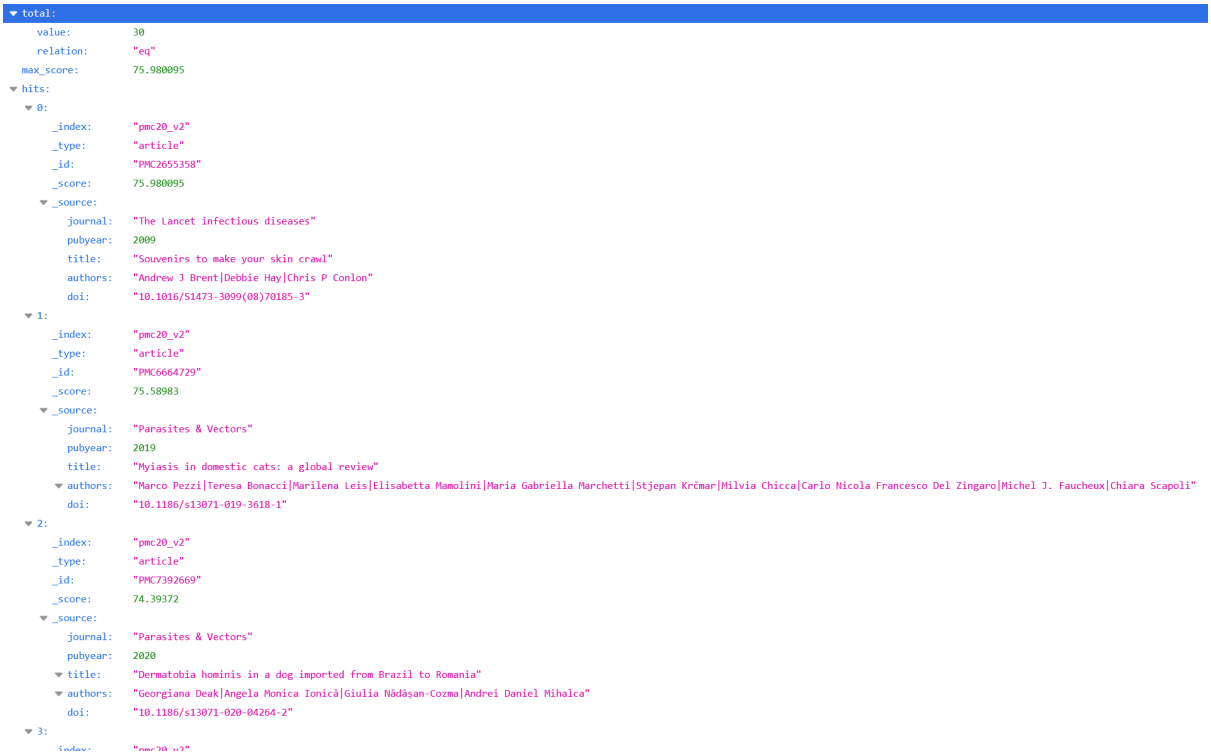

Fig. 13: *Result from the SIBiLS service for Bos taurus.*

The process can be repeated for the remaining species. As a result of our initial query, we thus have a list of species for which matches of the triplets have been identified in the literature. At the time of the writing, four species answer the question according to SIBiLS (see the appendix A for the details):

- *Bos taurus*
- *Homo sapiens*

- *Proteus mirabilis*
- *Canis familiaris*

Different nomenclatures reference the same concepts

For example, a taxon can reference the species, the genus, NCBI ID, Open Tree of Life, GBIF. When the query plan requires to chain two API calls, even if the input and output describe the same concept, a mapping between two nomenclatures might be required.

There is not always one-to-one correspondence between the two nomenclatures. For example:

- The NCBI ID of the species « Tyrannosaurus rex» is 436495.
- The NCBI ID of the species «Gossia ouazangouensis» doesn't exist.
- The NCBI ID of the genus «Gossia» is 375231.

5. Conclusion

The benchmark is showing a large set of questions and answers. A subset of questions have been picked up to elaborate the architecture of the FAIR Data Place, resulting in a few comprehensive workflow diagrams, which will serve to establish the technical basis of the FAIR Data Place. The ability to answer the rich set of collected questions will also depend on how powerful the federated endpoints API are. Given the relatively high heterogeneity across APIs (REST, GraphDB, SPARQL, FTP, ...) listed in the benchmark, a significant effort will be given to wrapping these API. This is clearly a dependency for the search and answering effectiveness of the FAIR Data Place. However the use case #2 clearly shows the feasibility of such an approach to answer complex questions. Further, the future development of WP11 will prioritise a subset of complex questions, with a special effort given to biotic interactions (see T11.2 with the pollinator use case), from which different services will be implemented.

6. Acknowledgements

The BiCIKL project is funded by the BiCIKL project (European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492). We would like to acknowledge all the contributors for the benchmark authoring effort and thank the BICIKL internal reviewers for the quality of their comments and suggestions.

7. References

PENEV, Lyubomir, KOUREAS, Dimitrios, GROOM, Quentin, *et al.* Biodiversity Community Integrated Knowledge Library (BiCIKL). *Research Ideas and Outcomes*, 2022, vol. 8, p. e81136. https://doi.org/10.3897/rio.8.e81136

VOORHEES, Ellen M., HARMAN, Donna K., *et al.* (ed.). *TREC: Experiment and evaluation in information retrieval*. Cambridge : MIT press, 2005. ISBN 9780262220736.

ABRAMI, Giuseppe, STOECKEL, Manuel, et MEHLER, Alexander. TextAnnotator: A UIMA based tool for the simultaneous and collaborative annotation of texts. In : *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020. p. 891-900. https://aclanthology.org/2020.lrec-1.112

WILKINSON, Mark D., DUMONTIER, Michel, AALBERSBERG, IJsbrand Jan, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 2016, vol. 3, no 1, p. 1-9. https://doi.org/10.1038/sdata.2016.18

8. Appendix

Appendix A

Factoid questions (e.g. Wh-questions such as where, when, who, ...): questions which can be answered with a concept of a short phrase

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
1.1	What species are predators of bats ?	Biotic interaction	Corallus hortanus	https://www.gbif.org/species/2464951	HOPKINS & HOPKINS (1982) mention an attack by an unidentified snake on a bat (probably Phyllostomus discolor Wagner, 1843) in the region of Manaus, in northern Brazil (the snake was described as an "arboreal constrictor", and was most probably a Corallus hortulanus (Linnaeus, 1758);	https://doi.org/10.1590/S0101-81752007000300036	Globi	
			C. hortulanus	https://www.gbif.org/species/2464951	MARTINS & OLIVEIRA (1998) found a bat (Myotis sp.) of 55 mm of total length and 7 g in the stomach of an adult C. hortulanus from the Rio Jaú region in Amazonian Brazil.	https://doi.org/10.1590/S0101-81752007000300036	GBIF	

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
			Zamenis longissimus		habitats of Bulgaria, high-lighting the importance of the Aesculapian snake (Zamenis longissimus) predation on bats in the Western Palearctic. Until now, 11 species of bats have been recorded as preys of snakes in Europe. Our observations are the first records of snake hunting on Mediterranean horseshoe bats (Rhinolophus euryale) and on greater mouse-eared bats (Myotis myotis) in Europe, and only the third to fourth observation of underground pred	https://doi.org/10.1515/mammalia-2018-0079	GBIF	
			Epicrates cenchria		This study describes the event of predation of an Epicrates cenchria on a Desmodus rotundus, in a cave in Tena, Ecuador.	http://dx.doi.org/10.3897/subtbiol.19.8731		
		Biotic interaction	Procyon lotor			https://link.springer.com/article/10.1007/s42991-020-00087-x		
1.2	What is the origin of SARS-Cov-2 ?	Geographic location	Wuhan, China	GPS 30.583332, 114.283333	The new decade of the 21st century (2020) started with the emergence of a novel coronavirus	https://europepmc.org/article/MED/32161092		

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
					known as SARS-CoV-2 that caused an epidemic of coronavirus disease (COVID-19) in Wuhan, China			
1.3	What diseases are transmitted by ticks ?	Biotic interaction / disease	Lyme borreliosis	ICD10:A69.2	Lyme disease, also known as Lyme borreliosis, is an infectious disease caused by the Borrelia bacterium which is spread by ticks	https://en.wikipedia.org/wiki/Lyme_disease		
1.4	What human genes are involved into Covid-19 infections ?	Genomics	ACE2	UNIPROT:Q9BYF1			UniProt	
1.5	What are the main Variants of Concern for SARS-Cov-2 ?	Genomics variants	N501Y		variants with the L452R or E484K substitution in the spike protein, the combination of K417N, E484K, and N501Y, or the combination of K417T, E484K, and N501Y substitutions in the spike protein	https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html		
1.6	What drugs have been active against SARS-CoV-2 in animal studies?	Drugs	GC-376	PMID 33953295	The treatment with GC-376 slightly improved survival from 0 to 20% in mice challenged with a high virus dose at 105 TCID50/mouse.	https://pubmed.ncbi.nlm.nih.gov/33953295/	MEDLINE	
1.7	What cell lines should be used to study sars-cov-2	Cell lines	Vero 6					

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
	?							
1.8	What is the current accepted name for all derivative components of specimen collection event Seigler 16161 on 27 May 2007. (For example (but not this specific collection event), a botanist collects three duplicates from a tree. These get sent to three collections and get differently curated. Data related to this collection event is now in three collections, ENA and cited in literature. A taxonomic expert reidentified a specimen (CoL) at one of the herbaria. What name is the one that was placed on the data by a taxonomic expert? How	taxonomic interaction	Mariosousa russelliana (Britton & Rose) Seigler & Ebinger	https://www.gbif.org/occurrence/1675983215/cluster	specimen reidentifed by known expert, correctly published			

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
	can that update be sent to the other components?)							
1.9	At which altitude was found the holotype Phragmataecia Newman, 1850 ?	Altitude	1100 m	http://treatment.plazi.org/id/0D5187AE-FFB1-5C0D-FF04-45C15AE1FCB5	Habitat and biology. Flight period: April –June. Altitude: 450–1100 m. Material examined: male (holotype), O. Afghanistan, prov. Nengrahar, D. Povolny (MWM); 1 female, O- Afghanistan, Sarobi, 1100 m, 28 .0 6.1956, Amsel leg. (ZSSM).	http://tb.plazi.org/GgServer/html/0D5187AEFFB15C0DFF0445C15AE1FCB5	Plazi	
1.10	Which collection can provide access to specimens of genus Latrodectus	Museum or Botanic gardens (~biobanks)	AMNH	???	??			
1.11	What are the parents of the hybrid "Solanum × michoacanum" ?	Parental linking	S. bulbocastanum × S. pinnatisectum		be a natural hybrid of S. bulbocastanum × S. pinnatisectum.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258389/	EuropePMC	Use case Hackathon
1.12	Where can I find a specimen of Dodo Raphus cucullatus ?	Specimen and taxonomic relationship	MacLeay Museum	grid.1013.3		https://www.euppublishing.com/doi/10.3366/anh.2013.0149	Cross-Ref	T11.2 & T11.4

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
1.13	Who discovered the species Dodo Raphus cucullatus ?	Literature citation	Heyndrick Dircksz Jolinck	DOI: 10.1080/08912960600639400	Heyndrick Dircksz Jolinck led one of these explorations (Moree 1998, 2001), and it was probably his account that described the Dodo for the first time	https://www.tandfonline.com/doi/pdf/10.1080/08912960600639400		T11.2 & T11.4
1.14	What articles describe species Dodo Raphus cucullatus ?	Literature citation		Check GRID ? ORCID ?				T11.2 & T11.4
1.15	What are the transmission routes of coronavirus?							
1.16	What is the phylogenetically closest living relative of the Dodo (Raphus cucullatus)?		Nicobar pigeon (Caloenas nicobarica)	KX902236	Soares et al (2016) studies the columbiform radiation and supports the close relation between the Nicobar pigeon (Caloenas nicobarica) and the dodo and its sister species Pezophaps solitaria.	http://europepmc.org/article/PMC/PMC5080718	EuropePMC / ENA	

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
1.17	Give me all papers where human ABL1(F359V) is mentioned in full-text tables	Fact extraction	List of PMIDs or DOI...	PMID/DOI			OpenBioDiv, SIBiLS, Plazi, ...	Open Biodiv competency questions and use cases (Pensoft): Give me all papers where a certain element [Taxon name, Person, Sequence, Collection/Institution Code, ..) is mentioned in a given section or suppl. data (Introduction, Methods, Caption, Table, Images, ...)

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
1.18	Give me all papers where human ABL1(F359V) is mentioned in supplementary data images	Fact extraction	List of PMIDs or DOI...	PMID/DOI			OpenBioDiv	Open Biodiv competency questions and use cases (Pensoft): Give me all papers where a certain element [Taxon name, Person, Sequence, Collection/Institution Code, ..) is mentioned in a given section or suppl. data (Introduction, Methods, Caption, Table, Images, ...)

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
1.19	Give me all papers where A.Y.42 variant of SARS-Cov-2 is mentioned in conclusion	Fact extraction	List of PMIDs or DOI...	PMID/DOI			OpenBioDiv	Open Biodiv compentency questions and use cases (Pensoft): Give me all papers where a certain element [Taxon name, Person, Sequence, Collection/Institution Code, ..) is mentioned in a given section or suppl. data (Introduction, Methods, Caption, Table, Images, ...)
1.20	What viruses are shared between bat x and y?	Fact extraction	List of bat species		a comparison of the lists of bats that include the viruses		SIBiLS, TB, BLR	
1.21	What is the geographic distribution of the bats that share the highest number (same species) of	Fact extraction	List of bats, viruses, distribution				SiBiLS, TB, BLR, GBIF	

QID	Question	Category	Answer(s)	Concept ID (optional incl. Accession Numbers)	Evidence statement	Document supporting the answers (e.g. DOI, URL, PMID, ...)		Other information
	viruses							
1.22	What species occurs in a given paper/abstract/section ?	Occurrences					OpenBioDiv	
1.23	What biotic interaction is used by Dermatobia Hominis to trigger myasis in humans ?	Fact extraction	Phoresis					

Open questions: causal question (why ?, how ?)

QID	Question	Category	Answers		Evidence statement	Document suppoting the answer (e.g. URL, PMID, ...)		Other information
2.1	How raccoon impact population size of bats in Europe ?	Biotic interaction	N/A			???		
2.2	What is the origin of SARS-Cov-2 ?	Etiology	N/A		The origin of the severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) virus causing the COVID-19 pandemic has not yet been fully determined	https://europepmc.org/article/MED/33105685		

2.3	As a Conservation Planner I want to cross-check species identification against reliably identified specimens so that create a checklist of species. For this, images, sequence data, georeferences, traits are needed.	resource management, biodiversity conservation	N/A				GBIF, ENA, COL, TreatmentBank, EOL traitbank	
2.4	As a Researcher, Scientist I want to search for trait information of a certain species and answer the question: How do species traits change based on changes in the environment due to global warming? For this, digitized collections, location, date, high-resolution images of specimens, trait information are needed.	climate-change impact, responses to climate change	N/A				GBIF, COL, TreatmentBank, EOL traitbank, Morphbank	
2.5	Which spiders do Sceliphron wasps predate?	biotic interactions	Araneus diadematus	https://www.inaturalist.org/observations/90978247				
2.6	How does the coronavirus respond to							

	changes in the weather ?							
2.7	What insects are hosted by a particular plant? Example of a use case for this information - will help benefit conservation of species by guiding the public to improve biodiversity in their gardens and guiding planting for conservation purposes	biotic interactions, biodiversity conservation	N/A			https://www.aucklandcouncil.govt.nz/environment/plants-animals/plant-for-your-ecosystem/Pages/plant-to-support-birds.aspx		
2.8	How can understanding the ecosystem that bats live in better prepare us for the next spill over?	biotic interactions, biodiversity conservation, resource management	N/A					
2.9	How many and what are the species known to science?		XX number of species					

Appendix B

The following is a screenshot of matrix showing NCBI species & papers (DOI and/or PMCID) for the use case #2.

DOI	Bos taurus (30)	Homo sapiens (75)	Proteus mirabilis (2)	Canis lupus (2)
DOI 10.1016/S1473-3099(08)70185-3	✔	✔		
DOI 10.1186/s13071-019-3618-1	✔	✔		✔
DOI 10.1186/s13071-020-04264-2	✔	✔		
DOI 10.1007/s00436-014-3906-9	✔	✔		
PMC 478413	✔	✔		
DOI 10.4103/JLP.JLP_18_17	✔	✔		
DOI 10.1371/journal.pntd.0007858	✔	✔		
DOI 10.1590/1984-0462/2021/39/2020105	✔	✔		
DOI 10.4172/2324-8599.1000106	✔	✔		
DOI 10.1155/2012/371498	✔	✔		
DOI 10.4103/0974-620X.57313	✔	✔		
DOI 10.4103/0974-777X.116874	✔	✔		
DOI 10.4103/0974-620X.48422	✔	✔		
DOI 10.4081/pr.2012.e34	✔	✔		
DOI 10.1136/vetreco-2014-000072	✔			
DOI 10.1673/031.011.0114	✔	✔		
DOI 10.1016/j.abd.2020.05.018	✔			
DOI 10.1016/j.abd.2019.12.001	✔	✔		
DOI 10.3201/eid1401.070163	✔	✔		

DOI 10.1186/s13071-016-1823-8	✔	✔
DOI 10.3390/ani11010065	✔	✔
DOI 10.7717/peerj.2598	✔	
DOI 10.1016/j.ijpara.2013.06.007	✔	✔
DOI 10.1051/parasite/2019026	✔	✔
DOI 10.1016/B978-0-12-384947-2.00770-4	✔	✔
DOI 10.1186/1756-3305-7-22	✔	✔
DOI 10.1186/s13071-021-04742-1	✔	✔
DOI 10.1007/978-981-13-7252-0_5	✔	✔
DOI 10.1016/B978-0-7020-3369-8.00005-7	✔	✔
DOI 10.1016/B978-0-7020-5317-7.00006-0	✔	
DOI 10.4103/tp.TP_65_19		✔
DOI 10.1177/2324709618801692		✔
DOI 10.1590/S1678-9946202062047		✔
DOI 10.4269/ajtmh.18-0262		✔
DOI 10.1016/j.eucr.2020.101410		✔
DOI 10.1186/1471-2482-4-5		✔
DOI 10.7759/cureus.11905		✔
DOI 10.1590/S1678-9946201961045		✔
PMC 622460		✔
DOI 10.4103/0019-5154.143539		✔
DOI 10.1016/j.idcr.2019.e00531		✔
PMC 306096		✔

DOI 10.4269/ajtmh.20-1531	✔
DOI 10.1111/ijd.14848	✔
DOI 10.3201/eid1712.111062	✔
PMC 893314	✔
DOI 10.4103/0974-777X.93763	✔
DOI 10.3201/eid1612.100938	✔
DOI 10.4103/0975-7406.114316	✔
PMC 385545	✔
DOI 10.1016/j.eucr.2020.101303	✔
DOI 10.1177/1179547619869009	✔
PMC 256061	✔
DOI 10.1590/S1677-5538.IBJU.2016.0084	✔
DOI 10.4103/0974-620X.142607	✔
DOI 10.7759/cureus.10617	✔
DOI 10.1016/j.ijwd.2019.04.022	✔
DOI 10.3347/kjp.2018.56.2.199	✔
DOI 10.3347/kjp.2017.55.3.327	✔
DOI 10.1155/2012/483431	✔
DOI 10.1099/jmmcr.0.005151	✔
DOI 10.7759/cureus.8585	✔
DOI 10.4103/idoj.IDOJ_589_20	✔
DOI 10.1016/j.tmaid.2017.10.003	✔
DOI 10.3390/plants9010033	✔
DOI 10.1016/B978-1-4557-4801-3.00324-6	✔

DOI 10.4103/idoj.IDOJ_559_20	✔		
DOI 10.1186/s12886-018-1003-z	✔		
DOI 10.1155/2012/382917	✔		
DOI 10.3390/pathogens10020238	✔		
DOI 10.1016/B978-0-323-54696-6.00007-0	✔		
DOI 10.4103/0301-4738.195590	✔		
DOI 10.1016/j.abd.2019.06.001	✔		
DOI 10.1186/s13023-020-01534-1	✔		
DOI 10.1186/1746-4269-5-27	✔		
DOI 10.1016/j.idc.2017.10.009	✔		
DOI 10.1016/S1773-035X(15)30318-X	✔		
DOI 10.1016/B978-2-294-76382-3.00012-7	✔	✔	
DOI 10.1016/B978-2-294-70867-1.00055-X	✔	✔	
DOI 10.1002/vms3.370			✔