BiCIKL

Appendix 2 to the Grant Proposal of BiCIKL. Use Cases

Described below are three use cases, that show how a service as described above, the new community would aim to combine the terms complexity and rapidness as to provide sound solutions at the level and time required by connecting, linking, embedding published scientific information, in this way allowing complex biological questions to be answered in a way that currently is either impossible or highly time consuming for life scientists across Europe.

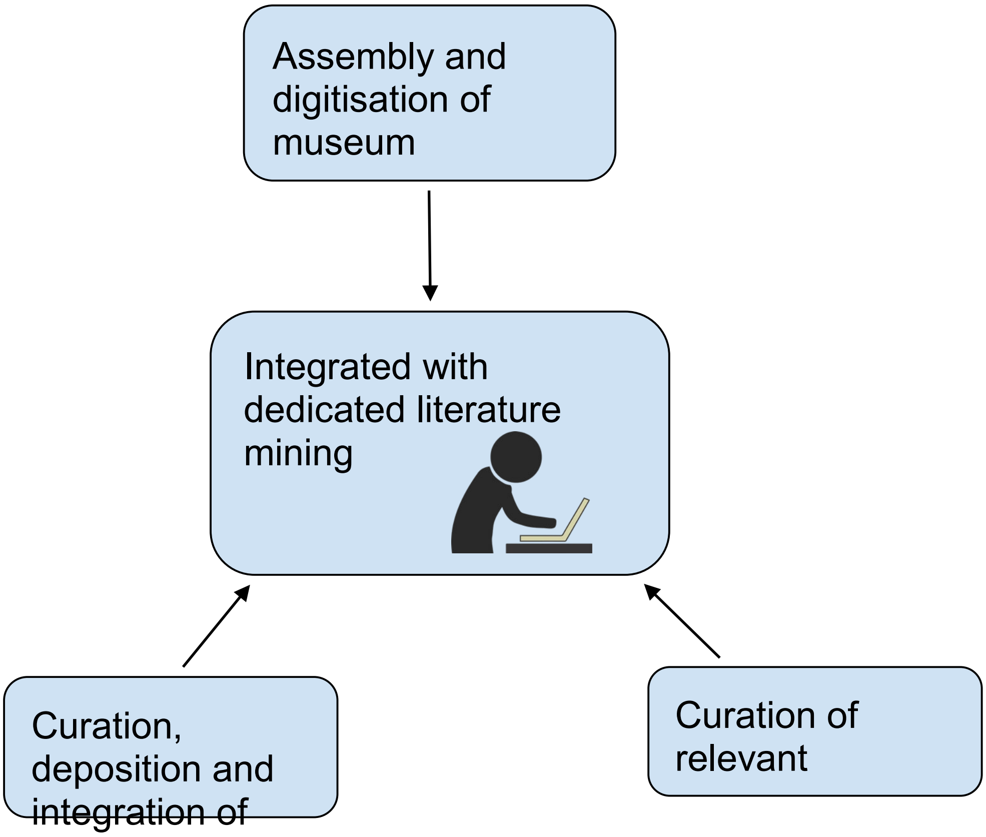
Assembly and digitisation of museum
Integrated with dedicated literature mining
Curation, deposition and integration of
Curation of relevant

Fig. 1. Exemplar use cases can explore research questions through various sets of interlinked data, produced in BiCIKL.

Exemplar Use Case #1:
A Rapid Response Workflow for Biological Invasions

Background

The rate of introduction of new exotic pests, invasive species and diseases is increasing, and is fueled by climate change, global trade and other changes to the environment (Seebens et al. 2017, Van Kleunen et al. 2015). Horizon scanning exercises, early warning and rapid response mechanisms can give us advanced warning of new emerging threats. However, it is clear that any lag between the emergence of these threats and dissemination of knowledge on these organisms will delay preventative action and increase the uncertainty under which decisions must be made.

For example, ash dieback disease was discovered in Poland in 1995, however, it took until 2014 before the correct name was given to it, *Hymenoscyphus fraxineus* (Baral, Queloz & Hosoya 2014). This inability to quickly connect data from a global corpus of literature, specimens and sequences could potentially make the difference between stopping a threat in its tracks and dealing with another major disease. It is a particular problem for invasive species where information may be available in countries where the species is native, but not accessible in countries where it is a problem. This issue is not restricted to diseases, but any alien organism is, almost by definition, dislocated from sources of information on it.

Imagine a new disease emerges on a tree crop in Europe. There is an urgent need to define a response. Should chemical control be used? Should infected trees be removed? Will the infection spread to other species? How is it being spread and how can we stop it? What other impacts might there be? To put effective policies in place urgent mobilization of evidence is needed and this needs to be assimilated together to provide actionable response options.

Proposal

To create a rapid delivery workflow that can be triggered at the first indication of a problem to deliver access to digitised biodiversity literature, taxonomic information, specimens and sequences for the selected taxon.

How could this be achieved?

1. Identification
    - Based upon genetic similarity, symptomatics and comparative morphology the pest species is either identified or matched to its nearest relatives.
    - The Catalogue of Life is used to specify the currently accepted name and synonyms of the taxon and its near relatives.
    - **Requirements:** Automated taxonomic and genetic workflows, based on APIs from the Elixir and DiSSCo infrastructures.
2. Extraction of traits from the literature
    - Using the Latin names from step one from the taxon and close relatives we will extract traits including habitat, species interactions and distribution.
    - With the taxonomic name(s), key publications are identified where this taxon is described.
    - Literature is digitized, read digitally and added to the pool of digitized literature in the Biodiversity Literature Repository.
    - Entity recognition is used against all treatments related to this taxon to extract, associated species, additional references, specimen references, physiological traits and morphological traits.
    - **Requirements:** Novel text-mining workflows to identify key components of taxonomic treatments.
3. Extraction of data from collections

   - Based upon the taxonomic names identified in step one we will identify specimens to image and database in the DiSSCo consortium.
   - These specimens, plus others that may have already been imaged, will be fully documented including georeferencing.
   - These specimens are combined with the data in step 2 to paint a complete picture of the organism, its relationships, biological interaction, morphology and physiology.
   - Access other observations and specimens globally to complement the European data
   - **Requirements:** Access to the imaging workflows of the DiSSCo consortium. Databasing and georeferencing pipelines. GBIF APIs to identify related observations and specimens. Aggregating workflows to assimilate knowledge from multiple sources.

4. Scenarios and Modelling
   - To inform decision-making processes models of potential future distributions will be created.
   - Options for responses will be evaluated against the evidence created in previous steps.
   - Future scenarios will be used to estimate the potential impact.
   - **Requirements:** LifeWatch modelling workflows to generate species distribution models based upon IPCC climate change scenarios.
5. Publication and Dissemination
   - A comprehensive summary of the results is published that links all the assimilated knowledge to the publications, sequence, specimens and observations that it was generated from.
   - Recommendations for response options, with detailed information on missing information and needs for future research.
   - **Requirements:** A publication pipeline that can allow researchers to work with the assimilated data and derived information seamlessly.

The Outputs

A rapid response mechanism that can discover and source literature and specimens; deliver digitized text and specimen images and create the links between specimens, literature, sequences and taxonomy.

References

Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme, PE, Jeschke JM, ... & Bacher S (2017) No saturation in the accumulation of alien species worldwide. Nature communications, 8: 14435. https://doi.org/10.1038/ncomms14435

Van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, ... & Antonova LA (2015) Global exchange and accumulation of non-native plants. Nature, 3;525(7567):100-3. https://doi.org/10.1038/nature14910

Exemplar Use Case #2:
Liberation and assembling of interlinked data on potential vectors of SARS-like viruses as a basis for meta analyses

Background

The COVID-19 pandemic has been caused by transmission of a virus from its host, most likely a bat, to a human with an ill-defined chain of interim vectors (e.g. snakes, pangolins). To understand the virus-host relationship, monitor it and prevent further outbreaks, the diversity and biology of bats and other hosts, have to be accessible rapidly and openly.

The scientific results from charting the world's biodiversity, that is the facts about species discovered by scientists, are kept in a corpus of hundreds of millions of pages. These include billions of facts, which are, in most cases, "imprisoned" by paywalls, copyright laws or just in hard-to-mine formats. Every day new data are added. For much of the world's species only one or a few articles exist, describing the species and adding some additional observations. Even for the well-known groups, such as birds and mammals, access to primary taxonomic literature (e.g. original descriptions) and reference works (e.g. faunas or catalogues) is almost impossible without extensive search, let alone that the data in these works are far from being accessible in structured and re-usable formats. For example, the reference work on bats cited in the Catalogue of Life (CoL), the Handbook of the Mammals of the World volume 9, is only accessible in print form. No link from the taxonomic name to any literature exists in this case in CoL. It literally takes days and a deep understanding of taxonomic literature to find existing sources, not to speak of making it available as FAIR data linked to specimens and sequences.

In a digital age, where research increasingly involves large international teams, and where there are tools for mining large corpora of texts, this situation needs to be changed. The COVID-19 pandemic is just one of the many examples where rapid **access to all possible data** is required. Usually, in such cases teams of specialists collate data in isolation from each other, which is doubtlessly an inefficient approach. On the other hand, anyone contributing to a large common data pool on the topic would help to build our broader understanding of pandemics, such as COVID-19. For example, there is already evidence for a possible link between the transmission of SARS-like Coronaviruses from their original hosts to humans. These hosts include a variety of wild animals, such as pangolins, bats, rats, snakes and civets. The evidence supporting these claims date back to the early 2000's (Li et al. 2005, Menachery et al. 2015) up to very recent papers published shortly after the Wuhan outbreak (Zhou et al. 2020, Lam et al. 2020). For example, Lam et al. (2020), provide a table including known viruses, strain and accession numbers allowing access to the respective data. The related taxonomic name of the host, however, is a digital dead end.

Nonetheless, and independent of the warning about possible outbreaks caused by Coronaviruses, no dedicated large-scale study on their potential vectors nor efforts for compilation and data mining of the published taxonomic and biological information available for these known reservoirs have been made.

For that reason, and in alignment with the recently announced COVID-19 Task Force[1] of the Consortium of European Taxonomic Facilities[2] (CETAF) (BiCIKL partner), the European Distributed System of Scientific Collections (DiSSCo[3]) consortium (RI partnering in BiCIKL) in collaboration with the similar collection network in the US, iDigBio[4], and related infrastructures, as GBIF and ELIXIR (RI partnering in BiCIKL), intending to create an efficient network of taxonomists, collection curators and other experts from around the globe, the BiCIKL partners Plazi[5], Pensoft[6] and SIB[7] launch an initiative to help make all taxonomic and other biological trait-related data about the hosts and vectors (e.g. pangolins, bats, snakes or civet cats) of the SARS-CoV-2 or other Coronaviruses

[1] https://cetaf.org/news/joint-cetaf-dissco-covid-19-task-force-call-contributions
[2] https://cetaf.org/
[3] https://www.dissco.eu/
[4] https://www.idigbio.org/
[5] http://plazi.org
[6] http://pensoft.net
[7] http://sib.swiss

accessible. In parallel, the SIB and the EBI are engaged in a customization effort to leverage the ELIXIR Data Platform with the aim to develop a literature triage system powered by a Covid-specific vocabulary (so-called COVoc[1]). The system shall not only support the curation of the peer-reviewed literature as available in PMC and MEDLINE, but it will also include bioRxiv and medRxiv pre-prints from the Allen AI corpus. The PMC, MEDLINE and Allen AI corpus are also directly available from the SIB Literature Services (SIBiLS) API.

Proposal

To harvest, mobilise and assemble interlinked data sets on potential hosts and transmission vectors of Coronaviruses, from various trusted resources: digitised biodiversity legacy literature, newly published semantically enriched papers, taxonomic information databases, data aggregators containing information on specimens and sequences.

How could this be achieved?

1. Creation of a bibliography on Coronavirus hosts, openly accessible on the Zotero platform with metadata enhanced with tags indicating the scientific names of the hosts.
Requirements: Access to publications corpora and tools to identify and retrieve publications with host records. Participation of the community to add missing publications.

2. Curation of bibliographic metadata to prepare this corpus of publications for batch-upload to a dedicated Coronavirus host community (*Coviho*) opened at Zenodo. Host names will be added as custom metadata fields, alongside the bibliographic references cited in the article.
Requirements: Refindit search & discovery services and Lycophron tool to upload to Zenodo. Custom metadata fields in Zenodo based on standard vocabularies to indicate host relationship of a taxon.

4. Use all hosts' taxon names to find the articles including treatments of the hosts, create a bibliography and add it to the Zotero public library (1).
Requirements: API to CoL to discover links from a taxonomic name to the cited article or taxonomic treatment, if possible including synonyms. Tools to resolve abbreviated references. Tools to find and download online publications. Service to scan and prepare print publications for text and data mining.

5. Harvest, extract and liberate taxon-specific data, including treatments, treatment citations, taxonomic names, figures, tables and bibliographic references. Annotate collection and specimen codes, and accession numbers with the respective persistent identifiers.
Requirements: Templates to prepare articles in different PDF styles. APIs to find and obtain PIDs for cited specimen, collection codes and sequence accession numbers. Tools for bi-directional linking. Workbench to correct manually data obtained automatically.

6. Convert the data extracted from these publications into FAIR data by uploading the treatments and figures to the Coviho repository at Zenodo, using the existing BLR workflow, adding their minted identifiers to the deposit of the original deposited publication on Coviho. Include supplementary file formats (e.g. BioC, DwC-A or RDF) that can easily be reused for further research.
Requirements: Converters to different data formats.

7. Launch a free-to-publish special journal issue in ZooKeys on potential hosts and vectors of Coronaviruses in a novel, semantically enhanced form, so that the published data could be submitted directly to and ingested by TreatmentBank and Coviho.
Requirements: Distribute information among the different communities of researchers to participate in the special issues; provide data auditing control on the quality and formats of published data.

8. Launch a free-to-publish special issue of non-conventional but valuable outcomes of Coronaviridae studies from various disciplines (for example, research ideas, grant proposals, methods, software, case

studies, single-media publications, etc.) in the open science journal Research Ideas and Outcomes[8] (RIO).
Requirements: Communication activities to convince researchers to publish early and non-conventional research outputs which will help establish open science practices, prevent duplication of efforts, increase international collaboration and help combating the COVID-19 pandemic.

9. Disseminate the publications as data sets to GBIF and other data aggregators where they are accessible with related observations and other records.
Requirements: Services to disseminate the data to target RI. APIs to integrate names of new taxa and synonyms to CoL+.

10. Implement AI and NLP algorithms to search for relationships between the taxa (viruses, hosts and vectors) to infer possible biotic or abiotic interactions that will shed light on the causes of epidemics.
Requirements: Services to download article or treatment sections of interest into a common document to provide NLP and AI methods. Search engines that could federate text corpora or Linked Open Data stores from different domains, e.g. Europe PMC (mostly biomedicine and molecular biology) and OpenBiodiv (biodiversity).

11. Raise awareness on the outcomes produced through existing operating scientific networks and environmental-related organizations as a source of FAIR, accurate and reliable information repository.
Requirements: Services to reach out to wide audiences. Specific communication campaigns linked to biodiversity sustainability on the long term .

The Outputs

Rapid and efficient access to interlinked data from various trusted biodiversity resources (databases, collections, biosamples and literature).

A test case on how semantic publishing can deliver data directly to the participating research infrastructures will be provided through publication of a dedicated topical journal issue.

The data sets assembled as a result from such a rapid text mining mechanism can be used for inferring new data-driven research hypotheses and knowledge (e.g. potential virus host relationships; distribution patterns of virus) and would serve as a working example for other multidisciplinary fields of science.

References

Li W, Shi Z, Yu M et al. (2005) Bats Are Natural Reservoirs of SARS-Like Coronaviruses. Science, 310 (5478), 676-679. https://doi.org/10.1126/science.1118391

Menachery VD,. Yount B, Debbink K et al. (2015) A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 21, 1508–1513 (2015). https://doi.org/10.1038/nm.3985

Zhou P, Yang X, Wang X et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273. https://doi.org/10.1038/s41586-020-2012-7

Lam TT, Shum MH, Zhu H. et al. (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature. https://doi.org/10.1038/s41586-020-2169-0

[8] http://riojournal.com

BiCIKL

Exemplar Use Case #3:
Cybercatalogue: Automating Cataloguing

Background

Torsten Dikow wants to write a taxonomic revision of the apiocerid flies. As a starting point, he needs a cyber catalogue of the existing taxa, their synonyms and augmentations with additional data through subsequent name usages. This will include taxonomic treatments - sections of scholarly publications - implicitly cited by the taxonomic name usage which represent the explicit history of the respective taxonomic names, referenced publications, cited materials, and gene sequences (Dikow & Agosti 2015).

With this catalogue in hand, he can expand the searches of existing services with all the possible combinations of taxonomic names, and thus discover gene sequences or specimens in collections identified by his predecessors under a no longer available name.

The TreatmentBank service will allow to process, in the best case fully automatically, the literature for Torsten and save an enormous amount of time. With his expertise he is the most qualified to curate the machine produced data which will become available as FAIR data in the Biodiversity Literature Repository, Catalogue of Life and GBIF. With this, any subsequent saves even more time by having a well curated corpus of data covering the apiocerid flies at hand.

Proposal

To provide a workflow that provides services to taxonomists and other interested parties to liberate and FAIRize the data from a target corpus of legacy publications.

How could this be achieved

In order to deliver, the proposed service will do the following

1. Assemble all the publications from the grantee, get the digital copies of the publications using related libraries; scan missing publications and create digital copies.
2. Import all the publications to BLR by curating the metadata including checking for existing DOIs of the articles that will be used if possible.
3. Prepare (if necessary OCR), and convert publications to the requested granularity, apply predefined standards
4. Make the data FAIR and open.
5. Link named entities to external resources (e.g. the accession code to the respective digital copy, treatment citations to the treatment and publication).
6. Disseminate the data to target services, e.g. GBIF taxonomic backbone.
7. Add the names to the project website that allows editing or an API .

Existing and required infrastructures are:

- Digitization: BHL
- Repositories: BLR/Zenodo
- Identifier minting: DOI via Zenodo, SIB
- Data Processing: SIB
- Conversion and TDM:
  - TreatmentBank
- Data access
  - Synospecies, Biodiversity Literature Repository and via API
- External resources

- ○ to link to specimens: DiSSCo, GBIF, natural history institutions
- ○ to link to genomic data: Mycobank, NCBI(-Europe)
- ○ to link to taxonomic names: Catalogue of Life+

The outputs

1. All publications: open if open access, open only for the consortium if closed access, but metadata are accessible and annotated in a deposit of the article in BLR.
2. FAIR data extracted from and linked to the source publications: Taxonomic treatment, figures, materials examined, named entities (taxonomic names, accession code, collection code. Data that are ready for test mining usage on BLR.
3. Cyber-catalogue as a service including all the treatments (i.e. synonymies, cited earlier mentioned treatments) in a form that can be reused (e.g. as RDF), other formats, and that can be edited and curated. The taxonomic names will be curated using the workbench in WP 9 and will be part c of GBIF taxonomic backbone.

This product is of general relevance, because through the expert curation the catalogue of taxonomic names including the synonyms, the links to external resources such as genomic data or specimens, and traits, can be used to combine datasets from different domains and different time of origines.

References

Dikow T, Agosti D (2015) Utilizing online resources for taxonomy: a cybercatalog of Afrotropical apiocerid flies (Insecta: Diptera: Apioceridae). Biodiversity Data Journal 3: e5707. https://doi.org/10.3897/BDJ.3.e5707

Miller JA, Braumuller Y, Kishor P, Shorthouse DP, Dimitrova M, Sautter G, Agosti A (2019) Mobilizing Data from Taxonomic Literature for an Iconic Species (Dinosauria, Theropoda, Tyrannosaurus rex). Biodiversity Information Science and Standards 3: e37078. https://doi.org/10.3897/biss.3.37078

Rivera-Quiroz F, Miller J (2019) Extracting Data from Legacy Taxonomic Literature: Applications for planning field work. Biodiversity Information Science and Standards 3: e37082. https://doi.org/10.3897/biss.3.37082

Exemplar Use Case #4:
Beyond cyber catalogues: What Linked Open Data can tell us on the systematics and evolutionary history of the centipede genus *Eupolybothrus* (Chilopoda: Lithobiomorpha)

Background

With a pace of some 18,000-20,000 new animal species described annually (http://www.organismnames.com) and the increasing number of collection specimens (half a billion just for insects), more than ever it has become necessary to harness the potential of the new technologies in genomics, collection digitization, and imaging in collection management (Short et al. 2018). On the other hand, the rate of species extinction has lent increased urgency to the description of new species and scientists have been forced towards a so-called ‘turbo taxonomy’ approach, where rapid species description is needed to manage conservation (Edmunds et al. 2013).

The centipede genus *Eupolybothrus* Verhoeff, 1907 comprises ca. 40 valid and doubtful species and subspecies. Eupolybothridans are known from the Eastern European and circum-Mediterranean countries, including large Mediterranean islands, Corsica, Crete, Cyprus, Sardinia and Sicily (Akkari et al. 2017, Stoev et al. 2010, 2013). It was among the first example groups for cybertaxonomic revisions which demonstrate new models in academic publishing (Akkari et al. 2017, Stoev et al. 2010, 2013). The innovative methods showcased already in 2010 (Stoev et al. 2010) include fine granularity XML tagging validated against the NLM DTD TaxPub for PubMedCentral, automatic dissemination of published content in XML format to various aggregators (GBIF, EOL, Wikipedia), vizualisation of all taxa via dynamically created Pensoft Taxon Profile (PTP) page, data publishing and, georeferencing of all localities via Google Earth. Furthermore, all suitable datasets were registered with ZooBank, GenBank and MorphBank.

Subsequently, a new holistic approach to taxonomic descriptions has been exemplified through the description of a new species of *Eupolybothrus* (Stoev et al. 2013). Thus, *E. cavernicolus*, a species discovered in a Croatian cave, has become the first eukaryotic species for which, in addition to the traditional morphological description, were provided a transcriptomic profile, DNA barcoding data, detailed anatomical X-ray microtomography (micro-CT), and a movie of the living specimen to document important traits of its behaviour. By employing micro-CT scanning, for the first time a high-resolution morphological and anatomical dataset was created, resulting in a ‘cybertype’ giving everyone virtual access to the specimen.

Proposal

Despite the recent progress on the taxonomy, genomics, ecology and imaging of the genus, several problems remain to be solved, for instance: 1) species being poorly described originally and currently being of uncertain taxonomic status; 2) species whose type material is unknown or lost; 3) species groups showing high genetic plasticity and/or cryptic species showing no variation in significant morphological traits; 4) lack of genomic data for most species; 5) lack of contemporary phylogenetic/phylogenomic analyses; 6) poor knowledge on the evolutionary biology and ecology of this taxon.

Here, we are going to develop the first complete taxon knowledge hub which will go beyond the notion of cyber-catalogues by pulling together all legacy and contemporary taxonomic, biological, and ecological data. Meta-analyses will be applied to link various types of data to reveal new patterns or test evolutionary hypotheses. We will also develop a conceptual framework, delivery and implementation workflow that can be used as a model for other taxa.

How could this be achieved?

1. Digitization of legacy literature and extraction of taxon treatments

- Legacy literature dealing with genus *Eupolybothrus* will be digitized and taxonomic treatments for all taxa will be extracted through GoldenGATE Document Editor[9].
- Taxon treatments will be processed and stored in the semantic knowledge graph OpenBiodiv. Through a system which utilizes semantic publishing workflows, text and data mining, common

[9] http://plazi.org/resources/treatmentbank/goldengate-editor

standards, ontology modelling and graph database technologies, OpenBiodiv establishes a robust infrastructure for managing biodiversity knowledge (Penev et al. 2019).

- In addition to OpenBiodiv all taxonomic treatments will be stored at Plazi and Zenodo.

2. Extraction of data from collections

- Major European museum collections will be investigated and specimens of *Eupolybothrus* will be examined, identified and properly catalogued.
- Available type material (holo-, paratypes, syntypes) will be established, described, photographed, standardized and catalogued.
- By employing micro-CT scanning, we will create a high-resolution morphological and anatomical collection of datasets and images for each species type individually that will allow virtual reconstruction – cybertype – according to Faulwetter et al.'s (2013) notion. For types which deem to be lost or not available for micro-CT scanning, we will create such virtual models based on non-type specimens. Micro-CT data will be stored at Morpho Source[10].

3. Answering evolutionary questions through OpenBiodiv

- Combining habitat and co-occurrence data we could establish biotic (feeding, preying upon, parasitizing) interactions with other organisms.
- Establish synonyms and enrich the existing pool of taxonomic descriptions for each individual species.
- Describe the geographic variation among the populations of model species.
- Link environmental factors (temperature, humidity, habitat, elevation) to morphological traits to establish unknown patterns.

4. Publication and dissemination

- A series of publications on the new concept, workflows and framework will be prepared and published in impact journals.
- The outputs of the OpenBiodiv, phylogenomic and taxonomic analyses will be published separately in relevant journals.
- Access to the cybertypes will be given through the Synthesys+ Virtual Access Facility[11] .

The Outputs

A first complete knowledge hub for a model taxon (*Eupolybothrus*) will be created to apply meta-analyses and establish new patterns or test evolutionary hypotheses. Furthermore, a new conceptual framework and a workflow will be suggested for testing with other taxa.

References

Akkari N, Komerički A, Weigand AM, Edgecombe GD, Stoev P (2017) A new cave centipede from Croatia, Eupolybothrus liburnicus sp. n., with notes on the subgenus Schizopolybothrus Verhoeff,

[10] https://www.morphosource.org
[11] https://www.synthesys.info/virtual-access.html

1934 (Chilopoda, Lithobiomorpha, Lithobiidae). ZooKeys 687: 11-43. https://doi.org/10.3897/zookeys.687.13844

Edmunds SC, Hunter CI, Smith V, Stoev P, Penev L (2013) Biodiversity research in the "big data" era: GigaScience and Pensoft work together to publish the most data-rich species description. GigaScience 2:14. https://doi.org/10.1186/2047-217X-2-14

Faulwetter S, Vasileiadou A, Kouratoras M, Dailianis T, Arvanitidis C (2013) Micro-computed tomography: Introducing new dimensions to taxonomy. ZooKeys 263: 1-45. https://doi.org/10.3897/zookeys.263.4261

Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K (2019) OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. Publications, 7, 38. https://www.mdpi.com/2304-6775/7/2/38

Stoev P, Akkari N, Zapparoli M, Porco D, Enghoff H, Edgecombe GD, Georgiev T, Penev L (2010) The centipede genus Eupolybothrus Verhoeff, 1907 (Chilopoda: Lithobiomorpha: Lithobiidae) in North Africa, a cybertaxonomic revision, with a key to all species in the genus and the first use of DNA barcoding for the group. ZooKeys 50: 29–77. https://doi.org/10.3897/zookeys.50.504

Stoev P, Komerički A, Akkari N, Liu S, Zhou X, Weigand A, Hostens J, Hunter C, Edmunds S, Porco D, Zapparoli M, Georgiev T, Mietchen D, Roberts D, Faulwetter S, Smith V, Penev L (2013) Eupolybothrus cavernicolus Komerički & Stoev sp. n.(Chilopoda: Lithobiomorpha: Lithobiidae): the first eukaryotic species description combining transcriptomic, DNA barcoding and micro-CT imaging data. Biodiversity Data Journal 1:e1013. https://doi.org/10.3897/BDJ.1.e1013

Exemplar Use Case #5:
Annotate Drosophilid genes using traits extracted from literature

Background

Drosophila melanogaster is an important model organism. The characterization of the fruit fly genome has numerous direct applications. They include the understanding of fundamental biological processes, such as aging, understanding human disease (*Drosophila melanogaster* is a reference model for many pathologies, including brain disorders), as well as research aiming to combat agricultural pests (e.g. *Drosophila suzukii*). Fruit flies are also used as an agent of biological control, meaning that an understanding of the reproduction process of flies is an important research subject with high economic impact.

Proposal

Judith Wagner is working at Salventa. She is interested in exploring the reproduction mechanisms of different Drosophilidae and sibling species (e.g. Tephritidae). She is especially interested in the ability of fruit flies to reproduce via parthenogenesis. Her investigations start with querying FlyBase. Unfortunately, she could not find any gene specifically associated with parthenogenesis. She therefore initiates a search in MEDLINE. After running several queries, which returned relatively outdated research, she realized that parthenogenesis is better described as “asexual reproduction” in the Gene

Ontology. Via GOA (Gene Ontology Annotation), she finally identifies a set of genes (Tramtrack-69, Mkrn1, …). She initiates a systematic search with all known genes associated with asexual reproduction and their orthologs. In parallel, she searches the literature to identify fly genera more likely to reproduce via parthenogenesis, and she uses Synospecies and Catalogue of Life to obtain all the taxonomic synonyms to broaden her search. .

How could this be achieved?

A broad coverage vocabulary to describe phenotypes of all species has been semi-automatically assembled, based on various terminological resources, including the Gene Ontology, Uberon, as well as the most common traits ranked by frequency in the taxonomic treatment section of Plazi.org. This last resource is key because many traits used in Plazi are not properly captured via the existing ontologies (e.g. gonopode related concepts). The whole MEDLINE has been pre-annotated with the terms, thus replacing all synonyms of a given concept by a unique identifier.

The Outputs

A ranked list of PubMed IDs has been identified to perform a more comprehensive curation effort on the subject.

Performance indicators

The improved engine should improve triage by about 80% compared to standard search.

Research activities

1. Define a simple controlled vocabulary based on taxonomic traits.
2. Annotate MEDLINE and PMC with this vocabulary via SIB Literature Services.
3. Index the annotations for effective and efficient search in EuropePMC.

References

EuropePMC curated content https://wellcomeopenresearch.org/articles/1-25

SIB annotation search and triage platform https://www.ncbi.nlm.nih.gov/pubmed/27374119

Example of application to crowd curated database: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz975/5622715