Open Call Project Report

21 October 2023

Donat Agosti
Pierre-Marie Allard
Deborah Caucheteur
Emmanuel Defossez
Tobias Kuhn
Tarcisio Mendes de Farias
Patrick Ruch
Adriano Rutz
Guido Sautter
Ana Claudia Sima
Felipe Simões
Philipp von Essen

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.

Table of contents

1. Open Call Project Overview

Title:
Exploiting and sharing existing and new knowledge in the frame of the Digital Botanical Gardens Initiative

Project Coordinators:
Pierre-Marie Allard, University of Fribourg, Switzerland

Project Members:
Emmanuel Defossez, Institute of Biology, University of Neuchâtel, Botanical garden of Neuchâtel, Switzerland
Adriano Rutz, Institute of Molecular Systems Biology, ETH Zürich, Switzerland
Ana Claudia Sima, SIB Swiss Institute of Bioinformatics, Switzerland
Tobias Kuhn, Knowledge Pixels, Switzerland / VU University Amsterdam, Netherlands
Tarcisio Mendes de Farias, SIB Swiss Institute of Bioinformatics, Switzerland

BiCIKL Research Infrastructures Involved:
SIBiLS, Plazi, EBI (ChEBI), CERN (BLR, Zenodo), OpenBioDiv

Non-BiCKL Research Infrastructures accessed:
NCBI (PubChem), Wikimedia foundation (Wikidata), Nanodash (Nanopublications)

Biodiversity data classes and services included:
Plants
Metabolites
Open Access and paywalled literature
Biodiversity
Chemodiversity
Open Science

Short background (1-2 sentences):
The “Exploiting and sharing existing and new knowledge in the frame of the Digital Botanical Gardens Initiative” project aims to harness BiCIKL research infrastructures to Improve past knowledge exploitation (extract pairs of chemical compound and taxon from literature) and Enhance future knowledge dissemination (design novels ways to share natural products occurrences).

Expected outcomes:
- A tabular file compiling referenced chemical structure - organism pairs extracted from literature. Should contain SMILES, IUPAC and/or common name of chemical structures; binomial denomination of organisms; and DOI of the reference documenting this occurrence. Additional metadata (material citation, collections, treatments, and specimen information) will complement the occurrence description. Will be contributed to LOTUS.
- A nanopublication (or alternative format) reporting the occurrence of metabolites evidenced in a computational mass spectrometry experiment.

2. BiCIKL Research Infrastructure

2.1 Results and Discussion

Circa 1000 pdf from the Phytochemistry journal have been processed through the Plazi pipeline and thousands of taxonomic treatments have been extracted and shared (see resume here). Several millions of LOTUS terms have been highlighted through the SIBILS pipeline and can now be observed at https://sibils.text-analytics.ch/. Furthermore, the very first Nanopublication template has been designed to report the experimental observation of a natural products occurrence (in this case exploiting the Experimental Natural Products Knowledge Graph dataset). An example of such Nanopublication supports the description of the occurrence of 21'-oxovobtusine in *Tabernaemontana coffeoides*.

September 2023		**Medline**	**PMC**	**PMC (Author manuscripts)**	**PMC (Supplementary data)**	**Plazi**	**Pensoft**
Nb documents (files for suppdata)		36,142,609	5,546,055	839,396	6,507,981	629,807	5,524
Nb annotations	**Lotus**	**31,056,431**	**53,110,654**	**10,458,980**	**2,978,799**	**47,551**	**5,507**
	PubChem (subset)	**338,414,792**	**993,787,345**	**189,916,224**	**76,175,901**	**7,836,483**	**749,143**
Avg biodiv chemicals anns/doc (anns/fils for suppD)		**10**	**189**	**12**	**239**	**13**	**137**

Table 1. Results of the literature annotation process using PubChem and Lotus identifiers according to different collections in SIBiLS.

2.2 Challenges and how did you overcome them

The alignment of nomenclature and vocabularies used across the different BiCIKL services is crucial and has been somehow overlooked.

PDF published in a format that can not be automatically processed because of line spacing. The spacing has been adjusted but might affect other processing. This is being tested.

We have started with regular meetings and set up a Github discussion thread to centralise progress and questions but the initial momentum was not kept. A more serious communication and coordination across project members could have clearly be beneficial.

2.3 Future work/sustainability of services provided

2.3.1 How can others access the services in the future?

For now these are not services but rather data extractions results. They can be accessed respectively on the SiBILS and Plazi platforms.

2.3.2 How will the services be sustained for future users?

Annotations of taxon and chemical compounds
The SIBiLS collections - enriched from Plazi's Taxonomic treatments - are receiving daily updates. They have been maintained for several years and the near future under the umbrella of the ELIXIR Data Platform. Further, the SBDe (Swiss BioData ecosystem), a research infrastructure directly funded by the SEFRI (Swiss State Secretary for Research and Innovation), is likely to support the services for up to 8 years.

3. Project Lead

3.1 Why did you choose to apply for BiCIKL funding? Which challenges/needs did you aim to address?

We are currently launching the Digital Botanical Gardens Initiative (DBGI, www.dbgi.org) with the ambition to explore innovative solutions for the acquisition, management and sharing of digital information acquired on living botanical collections. A particular focus is placed on the large-scale characterization of the chemodiversity of living plant collections through mass spectrometry. The acquired data will be structured, organised and connected with relevant metadata through semantic web technology. After validation and application in wild ecosystems, the gathered knowledge will inform ecosystem functioning research and orient biodiversity conservation projects.

One central aspect of the DBGI thus resides in the acquisition of high quality information on the digitised chemodiversity which in turns rely essentially in efficient metabolite annotation. For this we employ a taxonomically informed metabolite annotation process (https://doi.org/10.3389/fpls.2019.01329), which we have shown to systematically improve the performance of state-of-the computational metabolite annotation solutions. A requirement for this process is to access comprehensive ressources documenting the biological occurrence of small molecules. For this we recently opened the LOTUS ressource (https://elifesciences.org/articles/70780) to the community. However, numerous biological occurrences of natural products are still lacking.

We wanted to evaluate the BiCICKL Research Infrastructures to:

- *Improve past knowledge exploitation:* extract pairs of chemical compound and taxon from literature not present in LOTUS and relevant to the DBGI (papers describing phytochemical

investigation of plants of the Swiss Botanical Gardens, e.g. https://doi.org/10.1016/j.phytochem.2020.112469)

- *Enhance future knowledge dissemination:* explore publication solutions to efficiently disseminate knowledge acquired from metabolomics, that is, putative occurrences of chemical structures in biological matrices (simple RDF triples, Nanopublications (https://nanopub.org/) or similar).

3.2 Definition of the projects' success metrics in your own words.

Through the combined expertise of SIBILS, ARPHA, Plazi, BLR, GBIF and OpenBioDiv we expect to achieve the following outcomes.

For the "Improve past knowledge exploitation" part:

- A tabular file compiling referenced chemical structure - organism pairs extracted from literature. Should contain SMILES, IUPAC and/or common name of chemical structures; binomial denomination of organisms; and DOI of the reference documenting this occurrence. Additional metadata (material citation, collections, treatments, and specimen information) will complement the occurrence description. Will be contributed to LOTUS.

For the "Enhance future knowledge dissemination" part:

- A nanopublication (or alternative format) reporting the occurrence of metabolites evidenced in a computational mass spectrometry experiment.

In more general terms, the idea was to kickstart such exploration and connect with the relevant tools and resources creators of BiCIKL.

3.3 Were your expectations of the project outcomes met?

We have been able to advance on both stages of the proposed pilot. We have strengthened our connection with several actors of BiCIKL (Donat Agosti et al. at Plazi, Patrick Ruch et al. at SiBILS and Tobias Kühn et al. at Knowledge Pixels). So overall we can say that our expectations were met. However our initial goal on the "Improve past knowledge exploitation" will still require efforts and notably a better mining and curation of the terms connecting the chemical and biological sources.

3.4 What will you do with the results?

Of course these results only represent the first steps of a much bigger task. We hope that the initiated collaboration will continue and further steps have been identified. On the Plazi side, we hope to be able to digest other relevant corpus of literature (e.g. Journal of Natural Products). On the SiBILS side, we expect that the next iteration of the pipeline will link LOTUS terms using Wikidata ids. We also would like to build on the work done for the Biotic interaction ressource to highlight links between molecules and biological organisms (for example expanding the terms of the Relation Ontology). Regarding Nanopublications we aim

to build other relevant templates (for example, templates to describe an occurrence based on a publication and not from the ENPKG dataset).

4. Contribution to BiCIKL

4.1 How does this project contribute to the biodiversity community, possibly itemised by F.A.I.R dimensions?

Findability. All databases are freely available and they can be discovered by general web search engines. Further, SIBiLS is basically a search engine (together with an OpenAPI), which provides several exploration channels, e.g., keyword search, factoid question answering, SPARQL endpoints.

Accessibility. All data are publicly available with no restricted access but paywalled papers, whose only limited fragment can be made publicly available (e.g, meta-data and abstracts in MEDLINE).

Interoperability. Semantic interoperability is mediated via two main channels: usage of well documented standards and RDF. Non digital native articles are turned into JATS. For sake of annotations, articles are then transformed in BioC with entities marked up with RDF.

Reusability. All data generated by the participants, including WikiData contents, Plazi's treatments and SIBiLS annotations are available under a CC-BY 4.0 license.

4.2 Where can the community access the results/services provided?

All digitised contents are available on the web as well as in well identified resources, e.g. Zenodo, WikiData.

4.3 Provide links to outreach activities (if any).

None.

4.4 Recommendations for Future Open Calls.

Tbc

Annex

See example below provided by TreatmentBank (please delete all text in grey and example figure)

Results:
Processing and Scientific
Possible figures we could provide in a report as dashboards, see below for another project as example.
Number of publications (provided)
Number of publications (processed)

- With a Zenodo DOI
- Publication years

Percentage of Scanned x Digital-born
Percentage of Closed x Open Access
Number of pages (total)
Number of pages (processed)
Number of Bibliographic References
Number of Treatments
- In Zenodo/BLR
- In GBIF

Number of Treatment Citations
Number of New Taxa
Number of Materials Citations
- In Zenodo/BLR
- In GBIF
- In ENA
- In Matching Service

Number of Figures
- In Zenodo/BLR

Number of Tables
- In Zenodo/BLR

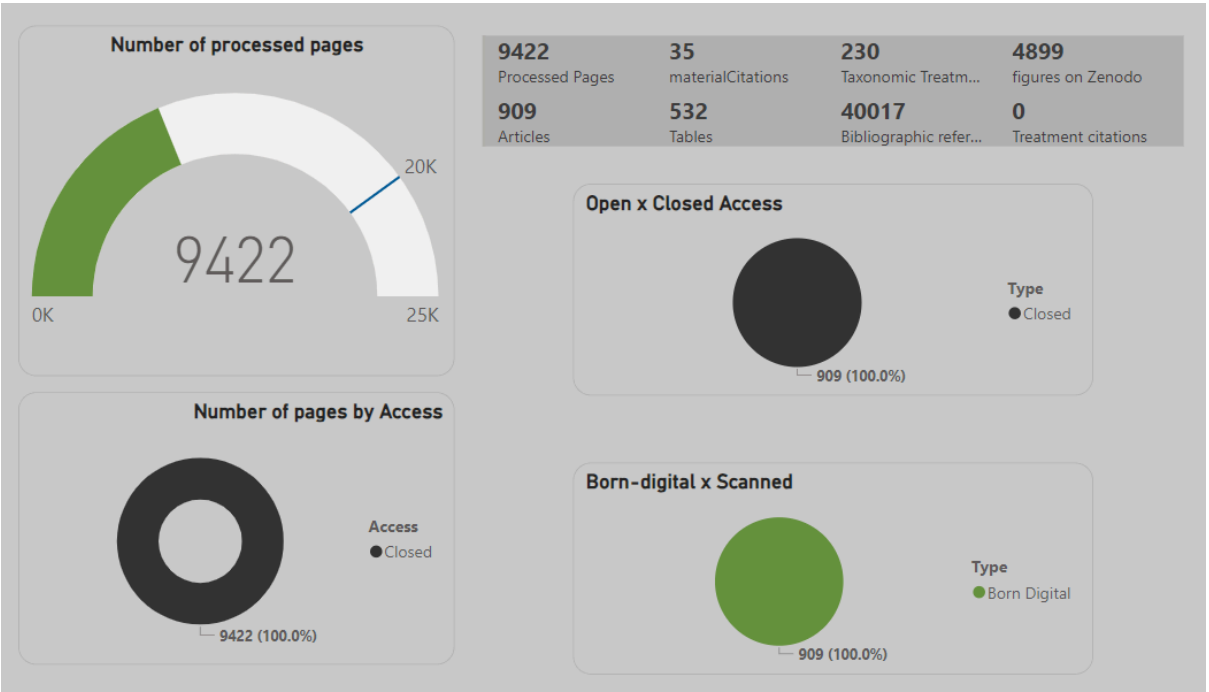

Fig Processing stats