Hacking Infrastructures Together: Towards better interoperability of infrastructures

Sofie Meeus[‡], Wouter Addink[§,|], Donat Agosti[¶], Christos Arvanitidis[#], Mariya Dimitrova[¤,«], Juan Miguel González-Aranda[#], Jörg Holetschek[»], Sharif Islam[§,|], Thomas S. Jeppesen[ˆ], Daniel Mietchen[ˇ,¦,ʔ], Tim Robertson[ʕ], Francisco Manuel Sanchez Cano[#], Maarten Trekels[‡], Quentin Groom[‡]

‡ Meise Botanic Garden, Meise, Belgium
§ Naturalis Biodiversity Center, Leiden, Netherlands
| Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
¶ Plazi, Bern, Switzerland
# LifeWatch ERIC, Seville, Spain
¤ Bulgarian Academy of Sciences, Sofia, Bulgaria
« Pensoft Publishers, Sofia, Bulgaria
» Botanic Garden & Botanical Museum Berlin-Dahlem, Berlin, Germany
ˆ Danish Natural History Museum, Copenhagen, Denmark
ˇ EvoMRI Communications, Jena, Germany
¦ University of Virginia, Charlottesville, United States of America
ʔ Data Science Institute, University of Virginia, Charlottesville, United States of America
ʕ Global Biodiversity Information Facility, Copenhagen, Denmark

Corresponding author: Quentin Groom (quentin.groom@plantentuinmeise.be)

Abstract

The BiCIKL Project is born from a vision that biodiversity data are most useful if they are viewed as a network of data that can be integrated and viewed from different starting points. BiCIKL's goal is to realize that vision by linking biodiversity data infrastructures, particularly for literature, molecular sequences, specimens, nomenclature and analytics. BiCIKL is an Open Science project creating Open FAIR data and services for the whole research community. BiCIKL intends to inspire novel, innovative, research and build services that can produce new and valuable knowledge, necessary for the protection of biodiversity and of our environment. BiCIKL will develop methods and workflows to harvest, link and access data extracted from literature. Yet, as the project gets underway, 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. To do this we organised a week-long hackathon where small teams worked on a number of pilot projects (Table 1) that were chosen to test the existing linkages between infrastructures and to extract novel ones.

We will present our experience of running a hackathon and our evaluation of how successfully it achieved its aims. We will also give examples of the projects we conducted and how successful they were. Finally we will give our preliminary evaluation of what we learned about the interoperability of infrastructures and what recommendations we can give to improve their interoperability, whether that is improvements to the data standards

©
.

used, the means to access the data and analyse them, or even the physical bandwidth and computational restrictions that limit the potential for research.

Keywords

molecular sequence data, nomenclature, specimens, biodiversity informatics, FAIR Data, Wikidata, literature, linked data

Presenting author

Quentin Groom

Presented at

TDWG 2021

Funding program

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

Grant title

Biodiversity Community Integrated Knowledge Library (BiCIKL)

Author contributions

SM and QG hosted the hackathon, had the initial concept and are the primary organizers. All authors provided research topics for the teams, led teams in the hackathon and contributed to the outcomes.

Conflicts of interest

The authors do not declare any conflict of interests.

Table 1.

Topics proposed for the BiCIKL Hackathon (September 2021) and the infrastructures and related organizations that will be linked. Details of each topic can be found on GitHub (https://github.com/pensoft/BiCIKL).

Title	Partners/infrastructures involved
Finding the lost parents	SIB, GBIF, CoL, Plazi, Wikidata
How good are Triple IDs in ENA?	ENA, GBIF
Enhance the GBIF clustering algorithms	GBIF
Assigning Latin scientific names to operational taxonomic units based on sequence clusters	UNITE/PlutoF, CoL
Registering biodiversity-related vocabulary as Wikidata lexemes and link their senses to Wikidata items	Wikidata, Plazi, CoL
FAIR Digital Object design from multiple sources	TDWG, GBIF, ENA, OpenBiodiv, DiSSCo
Enriching Wikidata with information from OpenBiodiv about type specimens in context from different literature sources	OpenBiodiv, Wikidata, GBIF, Plazi, Zoobank
Linking specimen with material citation and vice versa	SIB, BFH, MBG, GBIF, BGBM
Hidden women in science	Wikidata, Plazi, GBIF, CETAF, Science stories
An IPFS-Blockchain Interface	GBIF, Plazi, CETAF, Species 2000