Recommendations for interoperability among infrastructures

Sofie Meeus[‡], Wouter Addink[§,|], Donat Agosti[¶], Christos Arvanitidis[#], Bachir Balech[¤], Mathias Dillen[‡], Mariya Dimitrova[«,»], Juan Miguel González-Aranda[#], Jörg Holetschek[˄], Sharif Islam[§,˅], Thomas S. Jeppesen[¦], Daniel Mietchen[ˀ,ˁ,₵], Nicky Nicolson[ℓ], Lyubomir Penev[₰,₱], Tim Robertson[₳], Patrick Ruch[₴], 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
¤ Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy
« Bulgarian Academy of Sciences, Sofia, Bulgaria
» Pensoft Publishers, Sofia, Bulgaria
˄ Botanic Garden & Botanical Museum Berlin-Dahlem, Berlin, Germany
˅ DiSSCo, Leiden, Netherlands
¦ Danish Natural History Museum, Copenhagen, Denmark
ˀ EvoMRI Communications, Jena, Germany
ˁ Ronin Institute, Montclair, United States of America
₵ Institute for Globally Distributed Open Research and Education (IGDORE), Jena, Germany
ℓ Royal Botanic Gardens, Kew, London, United Kingdom
₰ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
₱ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria
₳ Global Biodiversity Information Facility, Copenhagen, Denmark
₴ Swiss Institute of Bioinformatics, Geneva, Switzerland

Corresponding author: Sofie Meeus (sofie.meeus@plantentuinmeise.be)

Abstract

The BiCIKL project is born from a vision that biodiversity data are most useful if they are presented as a network 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 make those links 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. In light of those aims, in the autumn of 2021, 74 people from the biodiversity data community engaged in a total of twelve hackathon topics with the aim to assess the current state of interoperability between infrastructures holding biodiversity data. These topics examined interoperability from several angles. Some were research subjects that required interoperability to get results, some examined modalities of access and the use and implementation of standards, while others tested technologies and workflows to improve linkage of different data types.

These topics and the issues in regard to interoperability uncovered by the hackathon participants inspired the formulation of the following recommendations for infrastructures related to (1) the use of data brokers, (2) building communities and trust, (3) cloud computing as a collaborative tool, (4) standards and (5) multiple modalities of access:

- If direct linking cannot be supported between infrastructures, explore using data brokers to store links
- Cooperate with open linkage brokers to provide a simple way to allow two-way links between infrastructures, without having to co-organize between many different organisations
- Facilitate and encourage the external reporting of issues related to their infrastructure and its interoperability.
- Facilitate and encourage requests for new features related to their infrastructure and its interoperability.
- Provide development roadmaps openly
- Provide a mechanism for anyone to ask for help
- Discuss issues in an open forum
- Provide cloud-based environments to allow external participants to contribute and test changes to features
- Consider the opportunities that cloud computing brings as a means to enable shared management of the infrastructure.
- Promote the sharing of knowledge around big data technologies amongst partners, using cloud computing as a training environment
- Invest in standards compliance and work with standards organisations to develop new, and extend existing standards
- Report on and review standards compliance within an infrastructure with metrics that give credit for work on standard compliance and development
- 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

Finally, the hackathons were an ideal meeting opportunity to build, diversify and extend the BiCIKL community further, and to ensure the alignment of the community with a common vision on how best to link data from specimens, samples, sequences, taxonomic names and taxonomic literature.

Keywords

hackathon, biodiversity informatics, API, Wikidata, cloud computing, data standards, FAIR data, linking

Preface

Providing services to science through data infrastructures is a complex and challenging job that requires juggling often conflicting needs of users, future developments, routine maintenance and software lifecycles. With all these pressures it is perhaps difficult to step back and evaluate where investment is needed and what the future opportunities are. This is one of the reasons that a hackathon was chosen as a mechanism to examine the interoperability of infrastructures (Suppl. materials 13, 14, 15). It allowed the people of infrastructures and their users to interact, somewhat separated from their daily routine and focus on just a single problem. BiCIKL is a highly technical project, however the route by which the technical challenges can be overcome is to enable relationships between people who want to work together. Each hackathon topic had its own aims and outcomes, many of which are being continued beyond the hackathon, yet, in this report we have tried to distil the problems of interoperability encountered by those projects. We intend to use these recommendations throughout BiCIKL to evaluate our progress towards better and longer lasting interoperability of biodiversity infrastructures.

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. 2021, Penev et al. 2022). Through working together, linking data, practising Open Science and Open Innovation, the project aims to make biodiversity data much more accessible and particularly to make them more interoperable with the ultimate vision of making biodiversity data more useful for novel research and informing policy decisions. In addition to the Open Science aspect of BiCIKL there are also the good practises 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 are not a prerequisite for complying with the principles, but do 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 (Fig. 1). Yet, both the technology and the community need to align with this vision, and hackathons can be a means to ensure this alignment.

Undoubtedly, the pandemic has presented a challenge to collaborative working, and particularly a hackathon that pre-pandemic was defined by the *radial collocation* of its participants (Pe-Than and Herbsleb 2019). Collocation enables participants to escape daily distractions and interruptions, focus on a single problem, but also exchange knowledge. Hackathons can expand someone's knowledge such that they can effectively plant the seeds of future innovation. Therefore, despite the challenges and risks associated with running and attending in-person events during the pandemic we believed it was worth the additional effort. Nevertheless, we are also aware that the travel restrictions imposed by the pandemic can limit inclusivity and so we organised the hackathon as a hybrid event.

A hackathon is an event of limited duration where teams tackle technical problems together, test ideas, create solutions, learn new skills, socialise and discuss (Angarita and Nolte 2020). A hackathon lacks the formality of a conference and is more hands-on than a workshop. It allows participants to escape from the limitations of their daily work, meet new people with different experiences and experiment with ideas and technologies they otherwise would not have the opportunity to do. Also unlike conferences and workshops they are specifically about collaboratively working towards technological solutions. Hackathons also can be the place to start collaborations in the long term and are an opportunity for professional development (Garcia et al. 2020). Hackathons can take a number of formats, but to describe ours we have applied the taxonomy of hackathons proposed by Kollwitz and Dinter (2019) (Fig. 2).

We also participated in the Biohackathon 2021. BioHackathons have been organised for almost twenty years to take advantage of the hackathon format in the life sciences (Garcia et al. 2020). In Europe the ELIXIR infrastructure has organised one for the past four years, including a virtual event in 2020 and a hybrid event in 2021.

Everyone from the BiCIKL community was encouraged to submit topics for pilot projects to test interoperability between the infrastructures. The topics were retrospectively grouped into three themes

1. research-based questions,
2. evaluating the infrastructures’ modalities of access, and the use and implementation of standards, and
3. testing technologies and workflows to improve linkage of different data types.

Below, we outline these topics and use them to support five high-level recommendations for infrastructures to improve their interoperability.

Recommendations to the infrastructures

1. Use of data brokers

In principle data infrastructures can be linked directly together. Stable identifiers of digital entities on one infrastructure can be maintained on another to link infrastructures in one direction, or there can be reciprocal links to traverse infrastructures in either direction. Indeed, such linkage is implied by the knowledge graph depicted in Fig. 1. Bi-directional linking implies that each cited infrastructure cites the citing infrastructure. For example, a specimen used in a taxonomic treatment should be cited in that treatment and at the same time the infrastructure holding the specimen should cite the treatment that cites the specimen. Bi-directional linking requires trust and coordination between infrastructures. Such close coordination is possible as demonstrated by GBIF and TreatmentBank, embedding Material citations and occurrence IDs respectively in their infrastructures (**topic 8**, Suppl. material 8). However, more often there is not sufficient incentive for two infrastructures to coordinate closely enough for bidirectional links to be supported.

An alternative to linking infrastructures is for a third party infrastructure to act as a broker between infrastructures. Wikidata is a collaboratively edited multilingual database hosted by the Wikimedia foundation (Vrandečić 2012), which can be used for this kind of data brokerage. Wikidata can be enriched in biodiversity data by the domain specific infrastructures, the community, but also other data brokers or knowledge graphs such as OpenBiodiv (**topic 7**, Suppl. material 7). The content can be managed manually on the website or through the API. **Topic 9** (Suppl. material 9) and **topic 11** (Suppl. material 11) used Wikidata in the hackathons as a broker to link together people, specimens and literature. Data brokerage is particularly important where multiple identifier systems exist, such as with person identifiers. ORCID identifiers can be used for living people who have opted to register, but Wikidata item IDs ("Q identifers") also act as a surrogate identifier for people (Van Veen 2019). Wikidata achieves this by consolidating the referenced resources in Wikidata into a single human entity type that is referenceable. No one single resource holds all the links between people, specimens and literature, also no one person identifier system works for every situation (Groom et al. 2020). In the hackathon, Wikidata was also used as a data broker for taxa. **Topic 12** (Suppl. material 12) used Wikidata as a bridge between GBIF and ENA for taxon IDs (based on NCBI taxonomy) and taxon names, because they use different taxonomic backbones that are joined within Wikidata. All these examples show that data brokers have a crucial role providing links between identifiers systems, creating links where there is no other source, and providing links that can be curated by the community.

There are several advantages of data brokerage through Wikidata in addition to direct linking. The broker infrastructure has an incentive to maintain the links, because that is a primary function of that infrastructure. Wikidata is open to editing from anyone, which both allows users to contribute and correct links, but it also means the people that need the links are incentivized to provide them. At first sight it seems that a data broker adds an additional point of failure and additional search and processing requirements. However, a data broker can link many infrastructures together simultaneously meaning that one additional broker system can join a whole family of infrastructures together. The main requirement is for infrastructures to keep their key identifiers stable, but there is clearly an incentive to maintain stable identifiers if those identifiers help link the infrastructure in both directions to a host of other data.

RECOMMENDATIONS

- If direct linking cannot be supported between infrastructures, explore using data brokers to store links.
- Cooperate with open linkage brokers to provide a simple way to allow two-way links between infrastructures, without having to co-organize between many different organisations.

2. Building communities/trust

BiCIKL is a project about building a community and trust between infrastructures is an essential aspect of interoperability that goes beyond the purely technical issues. If infrastructures are going to invest resources to interoperate with each other they need to know that the other infrastructures will use the systems and standards that are put in place; that they will be consulted on the design and implementation and that there will be sufficient stability that the interoperability will last, such as ensuring backwards compatibility.

The community, however, extends beyond the infrastructures to the users, whether they are data providers or downstream consumers of the infrastructure's services. The user community will not only make use of the linked infrastructures but will also contribute to it, for example, by enriching data brokers and providing user feedback to infrastructures. The infrastructures should facilitate the reporting of issues, including those issues related to incompatibilities between infrastructures. Good examples of issue tracking are in place, but need to be visible to the users and issues should be responded to promptly and constructively. GitHub is often used as an issue tracker and the ability to discuss, prioritise and label issues are important to building trust. Nevertheless, not everyone is comfortable using GitHub so if the infrastructure has a large number of non-informatics users then other forms of feedback and issue tracking might be necessary. Some infrastructures also provide a user forum where users can ask questions and debate issues. Such fora can be invaluable for providing support, self help and can be a place new features can be discussed. There are also many external fora where infrastructure services are discussed and it makes sense for these to be monitored by the infrastructures as a means to understand their community.

An important aspect to community building is that potential community members recognize other people in the community with common skills, needs and experience. So while preparing the hackathon we paid particular attention to the demographic and diversity of skills of the participants. For example, hackathons can tend to be biassed towards male participation (Briscoe 2014) and we believe the aims of the hackathon are best achieved through contributions from a broad coalition of researchers. To support this we ensured a wide range of topics, encouraging interaction across teams and encouraged leaders to collaborate (Richard et al. 2015). It is also worth noting that some infrastructures, such as Wikidata, actually give agency to their users to add data, make corrections and resolve their own problems with the infrastructure.

For example, **topic 5** (Suppl. material 5) developed a workflow to extract biodiversity-relevant terms from the literature and to convert them into Wikidata lexemes which - after a first check by experts - can be further edited by the community (Fig. 3). **Topic 9** (Suppl. material 9) also highlighted the importance of a volunteer community of (non-technical) experts to help out the scientific community in enriching the information on, in this case, people through suitable platforms such as the Wikimedia products and Bionomia.

Having an Open Source code-base might be another way that users could resolve their own issues within the community. All of the above build trust between infrastructures and between infrastructures and users. This builds engagement, avoids infrastructure being reinvented, supports both technical and social innovation, and is inclusive.

Technology can also be used to underpin trust in infrastructures (De Smedt et al. 2020). For example, **topic 10** (Suppl. material 10) investigated the possibility of using blockchain to encrypt data and track its provenance. This technology could be used to increase the trustworthiness of data, because the transaction ledger cannot be tampered with.

RECOMMENDATIONS

- Facilitate and encourage the external reporting of issues related to their infrastructure and its interoperability.
- Facilitate and encourage requests for new features related to their infrastructure and its interoperability.
- Provide development roadmaps openly.

3. Cloud computing as a collaborative tool

Cloud Computing technology provides the means for system developers to purchase computation and storage resources for a period of time without the need to acquire or manage physical hardware. This can bring real benefit under some scenarios, such as the need for high computation capacity for short periods of time, to scale a system up with growing demand or performing tests using different hardware configurations. The growing maturity of cloud computing services available, such as from Amazon and Microsoft now provide easy to use tools that enable a small team to quickly manage complex environments. Having access to this capability, along with recipes and tutorials for managing aspects like security and backup is an attractive proposition for any team.

An important aspect of cloud computing that is attractive to the BiCIKL project is the ability to collaborate. The infrastructures connected to BiCIKL are typically operated on an institutional network with limited possibility for external collaborators to get involved. Even though the software is often developed in an open source manner, it can be near impossible for an external person to reproduce the environment and contribute significantly. During the BiCIKL hackathon a portion of the GBIF infrastructure was recreated on the Microsoft Azure cloud for **topic 3** (Suppl. material 3) and access given to all participants. Following a brief introduction, participants were able to run routines on the shared environment, contribute code to GitHub and really collaborate around shared problems. Once tested on the shared space, the changes were brought into the production system at GBIF. This workflow demonstrated the ability to collaborate openly across institutions using shared infrastructure.

Beyond collaboration, cloud infrastructures also commonly offer various services built on massive-scale Machine Learning implementations. This includes powerful enrichment services such as georeferencing, computer vision, translating and data clustering. Infrastructures may make use of such state-of-the-art services to enrich the data they serve and make links to other infrastructures, benefitting from a scaling effectiveness they could not meet on their own. An example is handwritten text recognition for sparse and high variance text lines, such as occur regularly on scanned labels (**topic 12**, Suppl. material 12). Such tasks can strongly benefit from generic computer vision algorithms trained on large-scale datasets.

Importantly, it should be noted that cloud computing comes at a financial cost, which may be offset through grants offering free credit. The costs of operating the Azure cloud for this hackathon was funded through a grant from the Microsoft Planetary Computer programme. Computer Vision-based linking approaches were piloted on voucher credit, but could be quite costly if implemented on a larger scale.

RECOMMENDATIONS

- Provide cloud-based environments to allow external participants to contribute and test changes to features.
- Consider the opportunities that cloud computing brings as a means to enable shared management of the infrastructure.
- Promote the sharing of knowledge around big data technologies amongst partners, using cloud computing as a training environment.

4. Standards

It is a fairly obvious statement that adoption and continued compliance with community standards is a positive step towards interoperability (cf. FAIR principles; Wilkinson et al. 2016). Standards include the use of common terms, controlled vocabularies and also data models. Standards are not, and should not be, static instruments of interoperability. They provide meaning and structure to data, but they also influence the types and resolution of the data collected. Therefore, they are not independent of the intended uses of data, which leads to some of the disparities between competing standards and incomparable implementations of common standards. In cases where a small community is trying to connect with a larger one, adoption of the larger community's standards is a good first step. For example, the use of IIIF in **topic 9** (Suppl. material 9) immediately ensures interoperability with a large group of users. Yet, things do not always workout so smoothly.

As a case where standards are failing, **topic 1** (Suppl. material 1), focused on the standards regarding names of hybrids encompassed in the International Code of Nomenclature for algae, fungi, and plants (ICN). The ICN has recommendations for how to write the name of a hybrid, though the equivalent Code for zoology does not even make recommendations. The ICN's recommendations are not rules and are frequently not followed, as we discovered during the hackathon. Theremore, the ICN gives a lot of latitude to users for interpretation. When standards get used with real data, users discover their limitations and there has to be means for standards to accept feedback and evolve. A particularly thorny case of where the proposed standards have so far failed to survive real world implementations are that of identifiers for specimens. A single stable identifier for collection objects has long been seen as desirable and a challenge in the biodiversity informatics community (Guralnick et al. 2014). There have been many proposed schemes, such as LSIDs (Clark et al. 2004) and GUIDs (Nelson et al. 2018), yet no single system has prevailed. The so-called ''Darwin Core Triplet'' was once a popular solution. The concept was to create a unique identifier from the combination of the institution code, collection code and the catalogue number. It was adopted by members of the International Nucleotide Sequence Database Collaboration (INSDC), such as ENA. Yet it has many deficiencies, both in its uniqueness and in the variability in the way it is implemented (Guralnick et al. 2014). Currently, although INSDC databases are one of the largest users of this standard, it is of little use in automatically connecting specimens, and the need to accommodate other standard identifiers is pressing (Groom et al. 2021). **Topic 2** (Suppl. material 2) focused on this aspect, because although the use of Darwin Core Triplets has been discredited we still have a large legacy of data that needs interpretation. The work on this topic highlighted the many problems of using these Triplets as identifiers and phasing out their usage seem appropriate, particularly as more unique and stable alternatives are available (Güntsch et al. 2017). The lack of a universal identifier for specimens is why **topic 8** (Suppl. material 8) chose to link material citations in literature to GBIF records, rather than directly to specimen catalogues. The addition of the new term 'MaterialCitation' in the Darwin Core standard allows linking of the two representations of the same physical specimen.

In the case of taxa and taxon names **topic 12** (Suppl. material 12) wanted to link taxon names to their taxonomic IDs and their gene annotation. It encountered issues related to the lack of standardisation and harmonization schemas across data sources. The results obtained from the hackathon demonstrated an important number of broken connections of the above categories that lead to data related to specimens being missed. A lack of standards, competing standards, or a lack of adoption of standards is the common problem.

Looking forward to the future of biodiversity standards, the FAIR Digital Object **topic 6** (Suppl. material 6) focused on creating standardised digital objects and validating them with a Shape Expressions (ShEx). Having the means to validate features of the data, such as data types, values, properties and constraints is a valuable tool to support standards compliance in different infrastructures, though it is notable that none of the other topics mentioned the use of schemas or Shape Expressions to validate data and we wonder how often these are actually used in practise by infrastructures.

Standards need to be developed by a broad community to be useful to that whole community. But standards development and compliance need investment by infrastructures. Although widespread standards compliance across infrastructures would significantly enhance interoperability there are limitations to how far standard compliance can go. The primary objectives of the infrastructure come first and standards compliance has to compete for resources with other priorities. Nevertheless, there is a risk that infrastructure managers fail to see the potential for new users and uses of the infrastructure, because without standards compliance these potential users and uses are blocked and are therefore invisible.

RECOMMENDATIONS

- Invest in standards compliance and work with standards organisations to develop new, and extend existing standards.
- Report on and review standards compliance within an infrastructure with metrics that give credit for work on standard compliance and development.

5. Modalities of access

The ways that researchers access data can have a large influence on what research is conducted and how easy it is for researchers to do what they want. BiCIKL infrastructures aim to provide Open Data to be used however the users want. They want to support innovative uses and novel applications, but also more prosaic uses for the data. The aim is to do more and better science in a timely manner. The modes by which data are accessed is an important consideration in reducing the barriers and friction to use of these data. They are also critical to what uses can be made of the data. We recommend that infrastructures provide as many different modalities of access as possible. Only by doing this will they give access to the data without limiting the uses that researchers can make of the data. We have distinguished four basic levels of access, all of which have use to the community. These are:

1. browsing the data via a web portal,
2. programmatic access via an API,
3. downloading data to be used locally and
4. personal requests for unique sets of data.

In the hackathon topics all of these modes were used (Fig. 4). However, within these categories there are some nuances and it should not be assumed that one mode of access can substitute for another. For example, full data dumps can sometimes be achieved through scraping of web portals or an API, but these are poor substitutes for a properly implemented download facility.

Portal Access

Web portal access to the data allows users to evaluate what data is available in an infrastructure, in what format and what the quality and structure is like. They also support simple information requests. They are usually the first point of contact a researcher has with an infrastructure and are therefore critical to supporting a longer relationship with that researcher. If web portal access is slow, confusing or incomplete it is likely that the potential user will either go elsewhere or create their own resources.

Application Programming Interfaces (APIs)

Web APIs provide simple programmatic access to data. They can be built into workflows and made completely automatic and repeatable, keeping the output up-to-date with the latest data in the infrastructure. Tools can be built upon them to connect and retrieve information from multiple infrastructures at once (Suppl. material 4) and they can be written in such a way that users can get access to the data without causing authentication and capacity problems for the infrastructure. Nevertheless, when researchers need access to large amounts of data or access to data in an unusual way, they may not be suitable. They can be too slow, unreliable or do not provide the right kind of access. To avoid excessive use of services providers often have to throttle availability to users and only a brief break in internet connectivity can stop excursion of a workflow. Users are very much at the mercy of the implementation and of how well it is documented. For these reasons users often resort to local instances of the data, which is why downloads are important.

Personal requested data

A feature of several hackathon topics was the use of data provided from an infrastructure through personal contact with one of the administrators. This was to circumvent the limitations of the modalities of access provided, such as where a public API or download facility is not provided, or those facilities do not provide access to all the data or the data are in an unsuitable format. Personally requested data are sometimes necessary, but they are also an indication that there is an unresolved demand for access from users. It is very useful to researchers if infrastructures can support them with bespoke requests, however they are also problematic from several stand points. Such requests may only be possible due to personal contacts of the researcher with those in the infrastructure. This does not allow a level playing field for research. It is an inefficient way to provide data and it does not support reproducibility and citation, because it is more difficult to track provenance.

Downloads

Data science often requires large amounts of data to be analysed and the only way to process these data efficiently is to create a local copy. Infrastructures should provide download access to all or part of the data so that it can be easily retrieved 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.

Acknowledgements

The authors thank all the participants and their respective organisations for their contributions in the hackathons (see participant list in Appendix). In particular, we would like to thank the five invitees to the BiCIKL hackathon at Meise Botanic Garden: Christine Driller, Marina Golivets, Rukaya Johaadien, Sarah Vincent and Sabine von Mering for their participation and valuable contributions.

Funding program

The BiCIKL hackathon and participation in the Biohackathon Europe was funded by the BiCIKL project (European Union’s Horizon 2020 Research and Innovation Action under grant agreement No 101007492), DiSSCo Flanders (Research Foundation – Flanders research infrastructure under grant number FWO I001721N) and the Elixir BioHackathon Europe. Microsoft provided Azure cloud computing credit through the Planetary Computer programme to support the collaboration in topic 3.

Mariya Dimitrova was supported with funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 764840.

Hosting institution

Meise Botanic Garden, Belgium

Conflicts of interest

References

- Angarita MAM, Nolte A (2020) What Do We Know About Hackathon Outcomes and How to Support Them? – A Systematic Literature Review. In: Nolte A, Alvarez C, Hishiyama R, Chounta I, Rodríguez-Triana M, Inoue T (Eds) Lecture Notes in Computer Science, 12324. International Conference on Collaboration Technologies and Social Computing.

Springer Collaboration Technologies and Social Computing, 50–64 pp. [ISBN 978-3-030-58157-2]. https://doi.org/10.1007/978-3-030-58157-2_4

- Briscoe G (2014) Digital innovation: The hackathon phenomenon. URL: http://qmro.qmul.ac.uk/xmlui/handle/123456789/11418
- Clark T, Martin S, Liefeld T (2004) Globally distributed object identification for biological knowledgebases. Briefings in bioinformatics 5 (1): 59-70. https://doi.org/10.1093/bib/5.1.59
- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2): 21. https://doi.org/10.3390/publications8020021
- Garcia L, Antezana E, Garcia A, Bolton E, Jimenez R (2020) Ten simple rules to run a successful BioHackathon. PLOS Computational Biology 16 (5): 1007808. https://doi.org/10.1371/journal.pcbi.1007808
- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E (2020) People are essential to linking biodiversity data. Database 2020 https://doi.org/10.1093/database/baaa072
- Groom QJ, Dillen M, Huybrechts P, Johaadien R, Kyriakopoulou N, Fernandez FJ, Trekels M, Wong WY (2021) Connecting molecular sequences to their voucher specimens. BioHackrXiv https://doi.org/10.37044/osf.io/93qf4
- Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith V, Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017: bax003. https://doi.org/10.1093/database/bax003
- Guralnick R, Conlin T, Deck J, Stucky B, Cellinese N (2014) The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current Identifier Practices. PLoS ONE 9 (12). https://doi.org/10.1371/journal.pone.0114069
- Kollwitz C, Dinter B (2019) What the Hack? – Towards a Taxonomy of Hackathons. In: Hildebrandt T, van Dongen B, Röglinger M, Mendling J (Eds) Lecture Notes in Computer Science, 11675. Business Process Management. BPM 2019. Springer, 354–369 pp. [ISBN 978-3-030-26619-6]. https://doi.org/10.1007/978-3-030-26619-6_23
- Nelson G, Sweeney P, Gilbert E (2018) Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. Applications in Plant Sciences 6 (2): 1027. https://doi.org/10.1002/aps3.1027
- Page R (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e8767
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Barov B, Hobern D, Banki O, Addink W, Kõljalg U, Ruch P, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J (2021) Towards Interlinked FAIR Biodiversity Knowledge: The BiCIKL perspective. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.74233
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e81136

- Pe-Than EPP, Herbsleb J (2019) Understanding Hackathons for Science: Collaboration, Affordances, and Outcomes. In: Taylor N, Christian-Lamb C, Martin M, Nardi B (Eds) Lecture Notes in Computer Science, 11420. International Conference on Information. Springer Information in Contemporary Society. iConference 2019, 27–37 pp. [ISBN 978-3-030-15742-5]. https://doi.org/10.1007/978-3-030-15742-5_3
- Richard G, Kafai Y, Adleberg B, Telhan O (2015) StitchFest: Diversifying a College Hackathon to Broaden Participation and Perceptions in Computing. *46th ACM Technical Symposium on Computer Science Education*, Kansas City Missouri, March 4 - 7, 2015. Association for Computing Machinery, New York, *114–119* pp. [ISBN 978-1-4503-2966-8]. https://doi.org/10.1145/2676723.2677310
- Van Veen T (2019) Wikidata. Information Technology and Libraries 38 (2): 72-81. https://doi.org/10.6017/ital.v38i2.10886
- Vrandečić D (2012) Wikidata: a new platform for collaborative data collection. 21st International Conference on World Wide Web, *Lyon, France*, *April 16-20, 2012*. Association for Computing Machinery, New York, 1063–1064 pp. [ISBN 9781450312301]. https://doi.org/10.1145/2187980.2188242
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

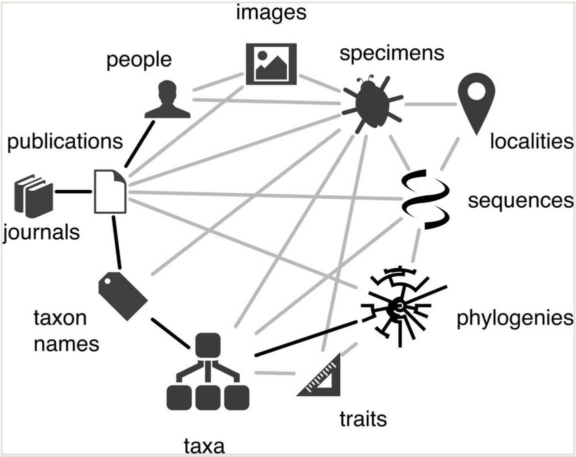
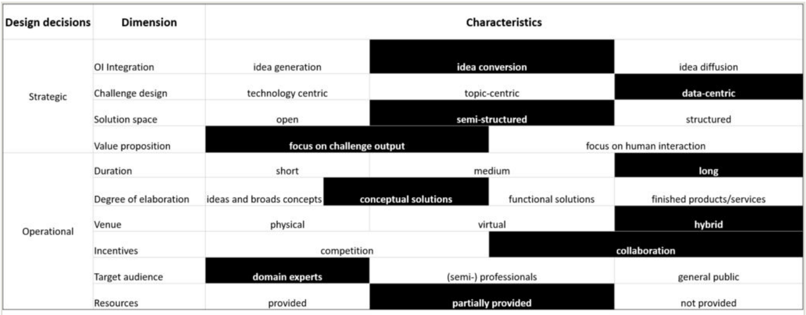


Figure 1.

A diagram of the biodiversity knowledge graph taken from Page (2016). This conceptual diagram shows the entities of knowledge on biodiversity and their linkages. However, even though these data are conceptualy linked it is not always possible to create actual links directly between infrastructures concerned with these different entities.



Design decisions	Dimension	Characteristics			
Strategic	OI Integration	idea generation	**idea conversion**	idea diffusion	
	Challenge design	technology centric	topic-centric	**data-centric**	
	Solution space	open	**semi-structured**	structured	
	Value proposition	**focus on challenge output**	focus on human interaction		
Operational	Duration	short	medium	**long**	
	Degree of elaboration	ideas and broads concepts	**conceptual solutions**	functional solutions	finished products/services
	Venue	physical	virtual	**hybrid**	
	Incentives	competition	**collaboration**		
	Target audience	**domain experts**	(semi-) professionals	general public	
	Resources	provided	**partially provided**	not provided	

Figure 2.

A description of the BiCIKL hackathon (black boxes) based upon the taxonomy of hackathons (Kollwitz and Dinter 2019). This gives an indication of how the BiCIKL hackathon was designed to achieve its aims.

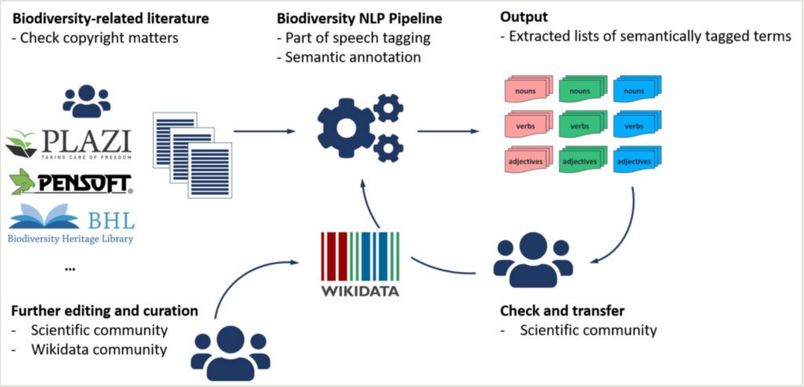
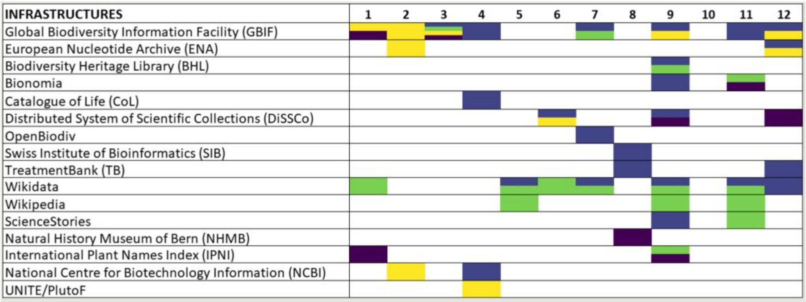


Figure 3.

A schematic workflow diagram of topic 5 showing the integration of multiple infrastructures and the user community in the process (Figure credit: Christine Driller).



INFRASTRUCTURES	1	2	3	4	5	6	7	8	9	10	11	12
Global Biodiversity Information Facility (GBIF)	🟨 🟪	🟨	🟦 🟩 🟨 🟪	🟦			🟦 🟩		🟦 🟨		🟦	🟦 🟨
European Nucleotide Archive (ENA)		🟨										🟦 🟨
Biodiversity Heritage Library (BHL)									🟦 🟩			
Bionomia									🟦		🟩 🟪	
Catalogue of Life (CoL)				🟦								
Distributed System of Scientific Collections (DiSSCo)						🟦 🟨			🟦 🟪			🟪
OpenBiodiv							🟦					
Swiss Institute of Bioinformatics (SIB)								🟦				
TreatmentBank (TB)								🟦				🟦
Wikidata	🟩				🟦 🟩	🟩	🟦 🟩		🟦 🟩		🟦 🟩	🟦
Wikipedia					🟩				🟩		🟩	
ScienceStories									🟦		🟩	
Natural History Museum of Bern (NHMB)								🟪				
International Plant Names Index (IPNI)	🟪								🟩 🟪			
National Centre for Biotechnology Information (NCBI)		🟨		🟦								
UNITE/PlutoF				🟨								

Figure 4.

The modes of access to the different infrastructures used by hackathon project teams: blue = Application Programming Interface or API (eg. SPARQL, RestFul); green = website, manual access; yellow = download or dump; and purple = personal request.

Supplementary materials

Suppl. material 1: Hackathon Topic 1

Authors: Quentin Groom
Data type: text
Brief description: Description of hackathon topic 1.
Download file (36.50 kb)

Suppl. material 2: Hackathon Topic 2

Authors: Jörg Holetschek
Data type: text
Brief description: Description of hackathon topic 2.
Download file (40.95 kb)

Suppl. material 3: Hackathon Topic 3

Authors: Tim Robertson
Data type: text
Brief description: Description of hackathon topic 3.
Download file (49.38 kb)

Suppl. material 4: Hackathon Topic 4

Authors: Thomas S. Jeppesen
Data type: text
Brief description: Description of hackathon topic 4.
Download file (43.89 kb)

Suppl. material 5: Hackathon Topic 5

Authors: Daniel Mietchen
Data type: text
Brief description: Description of hackathon topic 5.
Download file (49.57 kb)

Suppl. material 6: Hackathon Topic 6

Authors: Wouter Addink, Sharif Islam
Data type: text
Brief description: Description of hackathon topic 6.
Download file (43.89 kb)

Suppl. material 7: Hackathon Topic 7

Authors: Mariya Dimitrova, Lyubomir Penev
Data type: text
Brief description: Description of hackathon topic 7.

Download file (39.32 kb)

Suppl. material 8: Hackathon Topic 8

Authors: Donat Agosti
Data type: text
Brief description: Description of hackathon topic 8.
Download file (34.59 kb)

Suppl. material 9: Hackathon Topic 9

Authors: Maarten Trekels
Data type: text
Brief description: Description of hackathon topic 9.
Download file (46.28 kb)

Suppl. material 10: Hackathon Topic 10

Authors: Christos Arvanitidis, Juan Miguel González-Aranda
Data type: text
Brief description: Description of hackathon topic 10.
Download file (35.34 kb)

Suppl. material 11: Hackathon Topic 11

Authors: Quentin Groom
Data type: text
Brief description: Description of hackathon topic 11.
Download file (54.86 kb)

Suppl. material 12: Hackathon Topic 12

Authors: Mathias Dillen, Bachir Balech
Data type: text
Brief description: Description of hackathon topic 12.
Download file (35.93 kb)

Suppl. material 13: On-site participants of the BiCIKL hackathon at Meise Botanic Garden

Authors: Anja Van Ossel
Data type: Image
Brief description: Group picture of on-site participants of the BiCIKL hackathon at Meise Botanic Garden.
Download file (5.65 MB)

Suppl. material 14: Online participants of the BiCIKL hackathon at Meise Botanic Garden

Authors: Quentin Groom

Data type: Image

Brief description: Group picture of online participants of the BiCIKL hackathon at Meise Botanic Garden.

Download file (1.00 MB)

Suppl. material 15: List of participants involved in the BiCIKL hackathon and the BioHackathon Europe

Authors: Sofie Meeus

Data type: Table

Brief description: List of participants involved in the BiCIKL hackathon and participants in the BioHackathon Europe that worked on two biodiversity data related projects.

Download file (75.02 kb)