BiCIKL

Recommendations for interoperability among infrastructures

Deliverable D1.2

28 February 2022

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

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

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

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

[4] *Plazi, Bern, Switzerland*

[5] *LifeWatch ERIC, Seville, Spain*

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

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

[8] *Pensoft Publishers, Sofia, Bulgaria*

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

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

[11] *EvoMRI Communications, Jena, Germany*

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

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.

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

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

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

[16] *Swiss Institute of Bioinformatics, Geneva, Switzerland*

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Recommendations for interoperability among infrastructures
Deliverable n°:	D1.2
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP1
Lead beneficiary:	Meise Botanic Garden
Citation:	Meeus, S., Addink, W., Agosti, D., Arvanitidis, C., Balech, B., Dillen, M., Dimitrova, M., González-Aranda, J. M., Holetschek, J., Islam, S., Jeppesen, T. S., Mietchen, D., Nicolson, N., Penev, L., Robertson, T., Ruch, P., Trekels, M. & Groom, Q. (2022). Recommendations for interoperability among infrastructures. Deliverable D1.2 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 10
Actual submission date:	28 February 2022

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	16 February 2022	Meise Botanic Garden and authors
2.0	Review	23 February 2022	GBIF, ELIXIR Hub
3.0	Submission	28 February 2022	Meise Botanic Garden

Table of contents

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. 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 infrastructure.

Summary[1]

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 do so, we need to better understand the existing infrastructures, their limitations, the nature of the data they hold, the services they provide and particularly how they can interoperate.

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. Here, we give an overview of those topics, what their aims were, their methods, results and conclusions.

In addition, these topics and the issues in regard to interoperability uncovered by the hackathon participants inspired the formulation of following the 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) 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 reporting of issues and requests for new features related to their infrastructure and its interoperability.

[1] Modified from Meeus et al. (2021a).

- 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.
- 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 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.

In conclusion, 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.

1. Introduction

The overarching goal of BiCIKL is to create a community of infrastructures concerned with data on biodiversity through liberating data from scholarly publications and bi-directional linking of literature, taxonomic, DNA sequence and occurrence data (Penev et al. 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 these data more interoperable with the ultimate vision of making these 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 (Figure 1). Yet, both the technology and the community need to align with this vision, and hackathons can be a means to ensure this alignment.

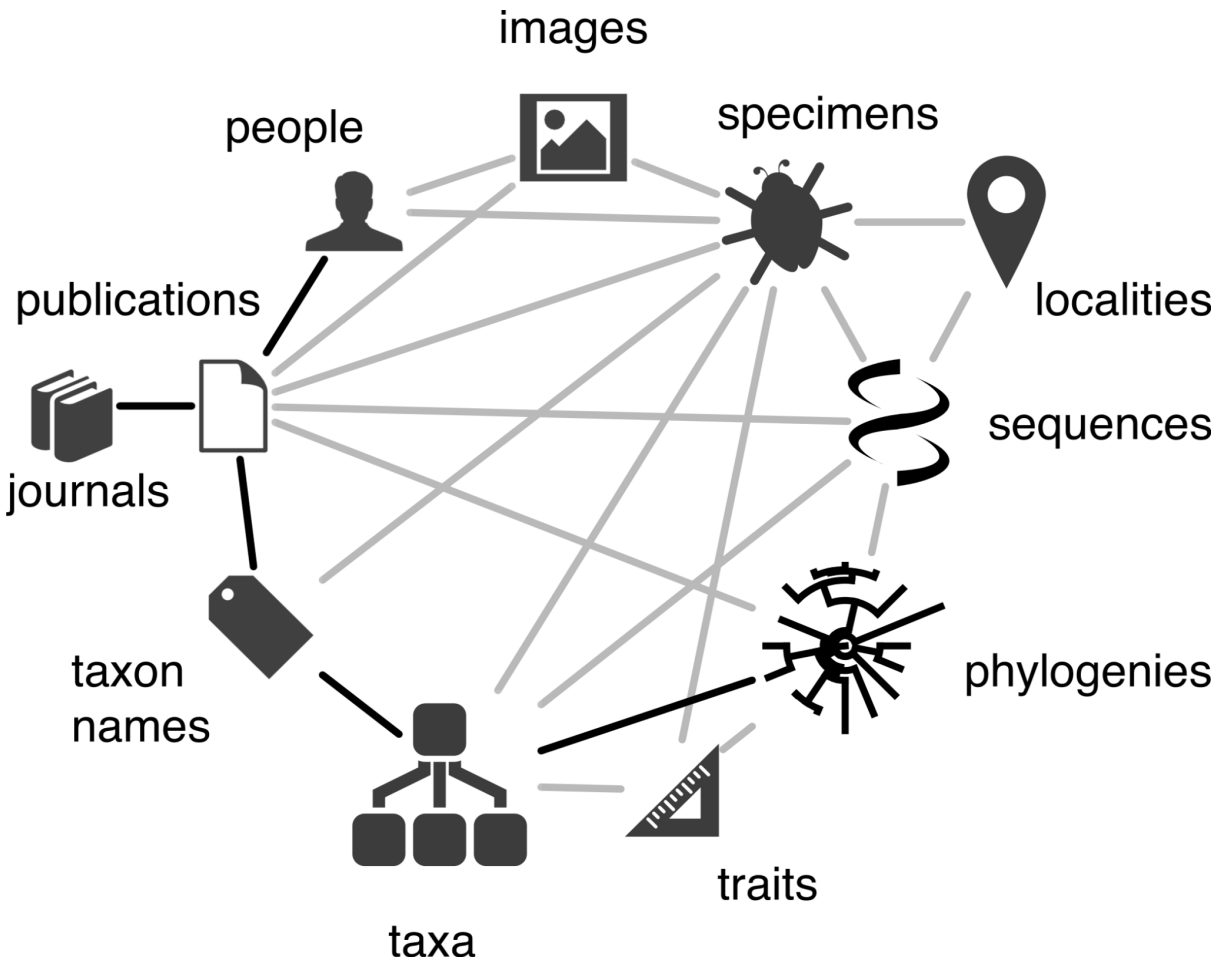

Figure 1. *A diagram of the biodiversity knowledge graph taken from Page (2016). This conceptual diagram shows the entities of knowledge on biodiversity and their linkages. However, even though these data are linked it is not always possible to create actual links directly between infrastructures concerned with these different entities.*

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 (Medina 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) (Figure 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 could be retrospectively grouped into three themes (i) research-based questions, (ii) evaluating the infrastructures' modalities of access, and the use and implementation of standards, and (iii) 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.

Design	Dimension	Characteristics		
Strategic	OI Integration	idea generation	**idea conversion**	idea diffusion
	Challenge desig	technology centric	topic-centric	**data-centric**
	Solution space	open	**semi-structured**	structured
	Value propositio	**focus on challenge output**		focus on human interaction
Operational	Duration	short	medium	**long**
	Degree of elabo	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.*

2. Recommendations to the infrastructures

2.1. 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 Figure 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**). 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**). The content can be managed manually on the website or through the API. **Topic 9** and **topic 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 numbers”) 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** used Wikidata as a bridge between GBIF and ENA for taxon IDs, because they use different systems 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.

2.2. Building communities and 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 (see 2.1) 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 the user to add data, make corrections and resolve their own problems with the infrastructure. For example, topic 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 (Figure 3). Topic 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 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.

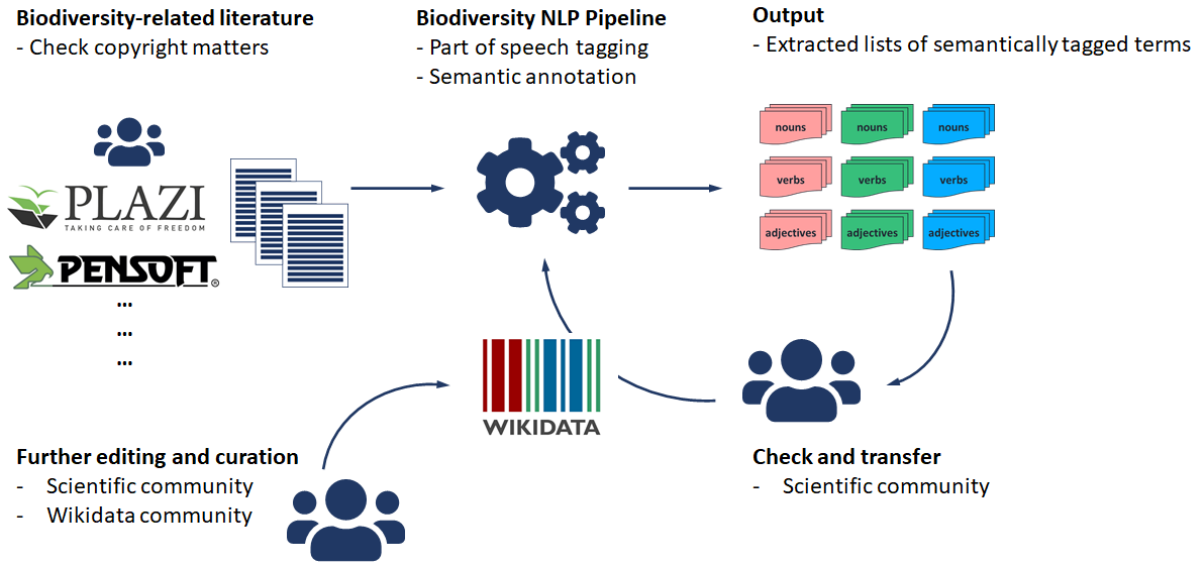

Figure 3. *A schematic workflow diagram of topic 5 showing the integration of multiple infrastructures and the user community in the process (Figure credit: Christine Driller).*

Recommendations

- Facilitate and encourage the reporting of issues and 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

2.3. 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** 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**). 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.
- 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.

2.4. 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** 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**, 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 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** 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** 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** wanted to link taxon names to their taxonomic IDs and their gene annotation. It encountered issues related to the lack of standardisation 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** 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 existing standards.
- Report on and review standards compliance within an infrastructure with metrics that give credit for work on standard compliance and development.

2.5. 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 (Figure 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 infrastructures. Tools can be built upon them 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 processed remotely by researchers. This could be provided in several ways. GBIF provides an asynchronous download system for queries and direct downloads of individual datasets. In the absence of a dedicated download system users may try to achieve the same result through an API, but this is highly inefficient for the user and infrastructure.

Recommendations

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

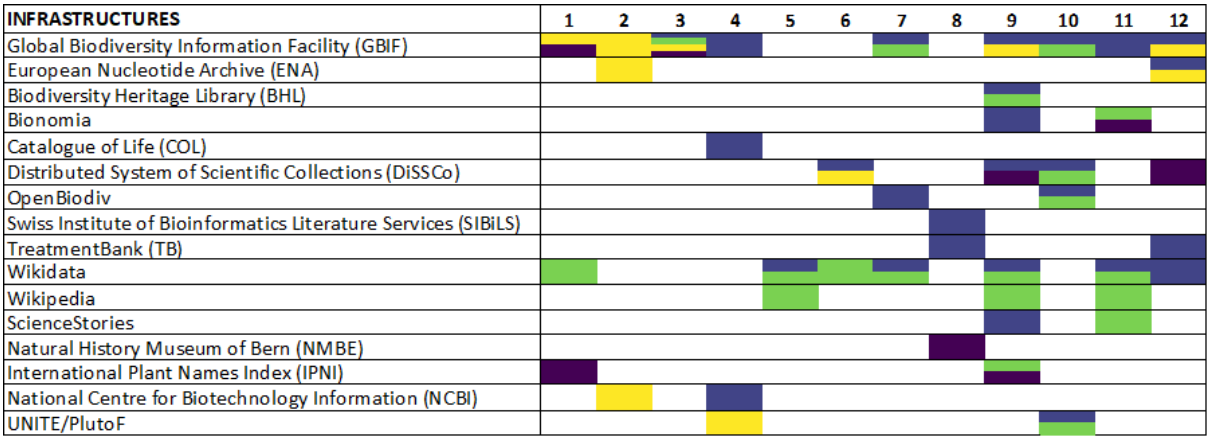

Figure 4. *The modes of access to the different infrastructures used by hackathon project teams: = Application Programming Interface or API (eg. SPARQL, RestFul); = website, manual access; = download or dump; and = personal request.*

3. Hackathon project summaries

3.1. Research-based topics

3.1.1. Finding the lost parents (Topic 1)

Aim/problem/goal

The idea for this topic resulted from research questions related to hybridization as a driver for plant speciation. For predicting the outcome (e.g. introgression, speciation, polyploidization) of plant hybridization it is important to know what the parental taxa are, and what their relatedness is. There is no single resource to discover what the parent species of a hybrid are. This is a particular problem in botanical research as a large proportion of plant taxa are of hybrid origin (Wissemann 2007). The goal was to compose an as long as possible, standardised checklist of hybrids and their parents taxa that can eventually be used for incorporation into other taxonomic resources, and to develop a workflow that automatically detects hybrids and their parents in publications.

Method

The parents of hybrids were parsed from literature, taxonomic checklists and Wikidata. This list was annotated with higher taxa hierarchies obtained from Global Biodiversity Information Facility (GBIF). Natural Language Processing tools - mostly rule based finite state automata - and named-entity recognition using gazetteer approaches were applied to deliver the annotations. A subset of articles from the Swiss Institute of Bioinformatics Literature Services (SIBiLS) were thus annotated.

Results

A checklist of 20,999 accepted hybrid names and a prototype of a tool for detecting hybrids and their associated parents from the literature.

Conclusion

Hybrids are ignored by Catalogue of Life and GBIF, yet during just one week of looking for hybrids and their parents, the team found a total of plant hybrids in the same order of magnitude as the number of species in the Asteraceae family (24,000), the largest family in the angiosperms. Knowledge about hybrids and their parentage is important to research in fields as diverse as evolutionary biology and the impacts of alien species. During the hackathon we have been linking literature and infrastructures such as GBIF to generate a list of hybrids and their parents (Fig. 4). By doing so, we came up with recommendations for extracting hybrid names from the literature for TreatmentBank, improving the documentation of three Darwin Core terms, and amending the International Code of Nomenclature for algae, fungi and plants to standardise hybrid names.

3.1.2. Assigning Latin scientific names to OTUs based on sequence clusters (Topic 4)

Aim/problem/goal

Curated sequence databases are important tools in modern taxonomy. They are used to identify sequences at the Operational Taxonomic Unit (OTU) level. OTUs are usually represented by some stable identifier, such as the Species Hypothesis (SH) in UNITE (Kõljalg et al. 2020) or the Barcode Index Number (BIN) in the Barcode of Life Data System (BOLD). In principle, these identifiers represent Species/Taxon concepts. In order to answer the question "What species does this sequence represent?" a linkage from an OTU identifier to a latin scientific name is needed, if existing. The taxon name for an OTU in the reference database has to somehow be derived from the taxonomic annotation of the sequences constituting the OTU. Currently, BOLD does not provide a single consensus taxon name for each BIN. In order to apply a taxon name, users therefore have to inspect all taxonomic annotations within the BIN and pick one. When blasting a single or a few sequences, this approach may suffice. However, in a taxonomic classification pipeline for many sequences (e.g. metabarcoding) this approach is impossible. Similarly, in order to place BINs or SHs into classic taxonomies such as the GBIF backbone taxonomy or the Catalogue of Life all bins must be unambiguously linked to a (parent) taxon. Therefore, the aim is to explore the algorithms for taxonomic assignment currently used by UNITE/PlutoF and the International Barcode of Life project (iBOL) Barcode Index Numbers (BINs) dataset in GBIF (The International Barcode of Life Consortium 2016) and discuss shortcomings and advantages. This project also aims to explore improvements based on the underlying data.

Method

A set of NCBI accessions with taxon labels assigned was used as input data. This could be imagined to either be the members of an OTU or top ‘X’ matches of a blast result set.

1. Clean/normalise names. i.e. informal names like ‘Bactrocera sp.27’ should be discarded at species level and be snapped to a higher taxon, in this example Bactrocera. This was best done using the GBIF species match API (https://www.gbif.org/developer/species#searching).
2. For all species level names, find the year of description and synonymy. Here, we used the Catalogue of Life (COL) nameusage search API (http://api.catalogueoflife.org/#/default/searchDataset).
3. For each accession with a species level taxon assigned, find out if the sequence was derived from type material. We accomplished this using a combination of the NCBI Entrez API ESearch and EFetch methods (Entrez Programming Utilities Help, https://www.ncbi.nlm.nih.gov/books/NBK25501/).

Results

During the hackathon we did "proof of concept" implementations of each of the three method steps in the R and Nodejs programming languages. Apart from further testing and improving error handling, the outstanding work would be chaining the steps into a pipeline that would

fulfil the objective of the topic. As an outcome of the work on step 3 i.e. retrieving type information from NCBI, we found that the NCBI Targeted Loci RefSeq projects are high quality data sources for Type specimens of Fungi and Prokaryotes. Hence a spin-off project in the form of an API adapter was written to make these projects available through GBIF (see Robbertse 2022 and McVeigh 2022) where they now contribute DNA sequences as well as bibliographic references to the clustered specimen view in GBIF.

Conclusion

APIs are already available to fulfil the goals of this topic (Fig. 4). However, these are spread across three different infrastructures and some subtasks require quite detailed knowledge of the underlying data structures. A full pipeline implementation of the proposed algorithm in for example R would therefore be a useful tool for taxonomic annotation of OTUs/sequence clusters.

3.1.1. Hidden women in science (Topic 9)

Aim/problem/goal

Due to various sociological and historical reasons, great achievements of women in science often disappear under the radar. It is essential to make sure that women are equally represented in the infrastructures and that their works are correctly linked to the person. The starting point of this project was to investigate the role of the infrastructures in improving the visibility of the 'hidden' scientists.

Method

The methodology of this project varied substantially from manual search of information inside the infrastructures and related infrastructures towards connecting APIs together to improve the visibility of the scientific achievements of women. Interconnectivity of the infrastructures is also what makes that scientific merit become more visible. It is also widely known that much of the achievements of people are locked-up inside natural history collections. Potential ways of liberating this information were investigated.

Results

During the hacking session, Wikipedia articles were jointly written, but also the profiles of over 200 women on Wikidata were added or completed. A big majority of these women also have a ScienceStories.io page available, showing also their specimens on Bionomia through IIIF compliant images. In order to engage more interested volunteers, a Wikipedia weekly session (Wikipedia Weekly 2021) was recorded explaining to a wider audience the power of connecting these tools together. Using GBIF data and an extract of the internal collection management system of Meise Botanic Garden, the potential of data enrichment was investigated.

Conclusion

In conclusion of this project, it is key to notice that a large community of highly motivated 'volunteers' is crucial in unlocking the information on hidden women, including suitable

collaborative platforms. APIs need to talk to each other, and linked open data needs to be used within the different infrastructures. Much of the information on these hidden women is either locked-up inside the natural history collections or inside of literature. It is therefore of most importance that this information is made available by the infrastructures, for example through DiSSCo, the Biodiversity Heritage Library (Kalfatovic et al. 2019), etc.

3.1.2. Join the dots: Making sense out of biodiversity data with a human focus (Topic 11)

Aim/problem/goal

Biodiversity data are often collected by teams of two or more people. But the names of the people in these teams are often just a string of characters on a label and not identifiable to people. If these people could be connected to their biographical data we could cross validate the data on labels and also examine the co-collecting profiles of collectors.

Method

We used data from Bionomia.net that links specimens on GBIF to people by their Wikidata Q numbers or ORCID. We also use Wikidata to extract data on gender and age of the collectors (Meeus et al. 2021b).

Results

A little more than 3000 co-collectors were joined in a network and a large number of the co-collectors formed a single large interconnected network. Women tend to have fewer co-collectors and female-female co-collectors only occurred from the 20th century onward. Co-collectors are most commonly close in age, but a wide range of age differences exists. Also about 5% of age differences are so large that we suspect that these are either errors in the biographical information or in the attribution of a person to a specimen.

Conclusion

Using Bionomia and Wikidata to analyse co-collection is a fairly simple process and is able to extract useful information. Nevertheless, this was only possible for people with a Wikidata Q number or to some extent people with an ORCID. However, this is only possible due to the retrospective assignment of identities to names on specimens. It would be far less prone to error if collectors registered for ORCIDs before depositing specimens in a collection, so that their identity was transparently recorded in the collection management systems from the time of deposition.

3.2. Topics evaluating the infrastructures' modalities of access, and the use and implementation of standards

3.2.1. How good are Triple IDs in ENA? (Topic 2)

Aim/problem/goal

Large sets of records in the European Nucleotide Archive (ENA) reference specimens (e.g. as specimen_voucher, bio_material, or culture_collection) through the use of GBIF triple IDs, that is a concatenation of institution code, collection code and catalog number. However, it is unclear whether these links correctly reference specimen records in GBIF as, for example, the collection code is sometimes omitted. The goal is to investigate how reliable the triple IDs are and to develop methods for improving them by inspecting additional data items (e.g. gathering date and country). Reliable links between ENA sequences and GBIF specimens would (a) allow users to follow links between both infrastructures, (b) feed metadata on voucher specimens from GBIF to ENA (which are often poorer on ENA), and (c) add publication references on ENA sequences to the corresponding GBIF specimens.

Method

The planned method for this task was to

1. Download ENA sequence records based on vouchers
   (referenced by specimen_voucher or bio_material or culture_collection)
2. Download/access potential specimens from GBIF
   (identified by inst/coll code and/or catalog number)
3. Develop matching algorithm
   - Triple ID based, if it exists
   - If not by examining metadata items (taxonomy, gathering date/country)
4. Result: List of (potential) matches (maybe with flag)

However, it soon became clear that these steps had already been done by GBIF. Shortly before the hackathon, GBIF had imported the sequences of interest from ENA. As all GBIF records, the clustering algorithm processed these records, already grouping 720k ENA sequences in clusters with their corresponding specimens. This team therefore joined the team that was working on enhancing the GBIF clustering algorithms (Topic 3) and worked on increasing this number by improving the algorithm.

Results and Conclusion

See summary topic 3

3.2.2. CAB2: A step towards Biodiversity data enrichment (Topic 12)

Aim/problem/goal

Natural history specimens may be sampled for sequencing, and these specimens/sequences published or cited in literature and deposited in repositories like ENA. Links between these

types of data are rarely explicit, so that it is not straightforward to connect a specimen to a sequence or literature. The goal of this project was to (re-)establish these links, making use of Computer Vision models and ad hoc text mining scripts to extract more data from the specimens.

Method

We made use of Computer Vision (CV) to identify indications of sequencing on specimen images. We then retrieved the corresponding sequences (in ENA), gene annotations and references in literature (in TreatmentBank and ENA flat files), either by matching identifiers or by mining through common properties such as taxonomic names and their corresponding identifiers. We also tried to leverage results from the GBIF clustering algorithms, which cover ENA sequences and published specimen data.

Results

A relatively small set of specimens was identified as (probably) having been sequenced (467 out of 3,184 specimens processed). There was considerable complementarity to previous results, but again only a fraction could be unambiguously matched to ENA sequences. Poor identifier propagation is a fundamental blocker, but we also had great difficulty in taxonomic matching between specimen and sequencing data, in particular through the ENA API. The GBIF clustering method was very conservative for this sort of matching and yielded almost no results, given the large variability and inconsistency in how identifiers are provided to the two infrastructures. The connection to literature showed similar issues with different representations of identifiers or even their total absence, with the added complication of trying to identify the material citations from text, tables or supplementary material in the first place.

Conclusion

Linking between these different data types is currently very difficult and labour-intensive. Scientists should be strongly encouraged to make use of persistent identifiers to maintain links in all sources and infrastructures to support and promote this. Computer Vision worked well, but showed scaling issues of costs. Sequencing labels are often a mix of sparse handwritten and typed text, for which free algorithms currently do not yield satisfactory results. In addition, the Computer Vision approach is likely to only cover specific cases and it's still challenging to connect the specimens to the sequences, as ENA identifiers are rarely used on the specimens themselves. Large-scale clustering approaches, such as performed by GBIF, could yield more results. Taxonomic interfacing between the different data sources should be improved. We made use of Wikidata as a broker, but these data are not always up-to-date and can suffer from taxon rank discrepancies. Taxonomic resolution options through the ENA API were very limited, so we had to resort to mining through data dumps instead.

3.3. Topics testing technologies and workflows to improve linkage of different data types

3.3.1. Enhance the GBIF clustering algorithms (Topic 3)

Aim/problem/goal

GBIF aggregates biodiversity data from many different sources such as citizen science platforms, specimen data from collections, literature and sequences. In 2020, GBIF developed a clustering algorithm to cluster these different types of data based on scientific name, location, date, etc. The aims of this topic were to explore enhancements to this previous work:

1. To build a common understanding of the process and software currently run by GBIF for detecting links across records and make improvements to this.
2. To explore the DataBricks platform as a tool for analysing and scripting processes that run across large data exports from GBIF.
3. To trial the use of a cloud environment to assess its suitability for collaboration across institutions.
4. To accommodate other research ideas from members in the hackathon.

Method

A databricks cluster was established on Microsoft Azure, using credits kindly donated by the Microsoft Planetary Computer programme to support our effort. An induction programme was run, presenting the DataBricks environment to the members of the group.

The team split into individual and group tracks and explored the following:

1. A data analysis of the EMBL's European Bioinformatics Institute (EMBL-EBI) datasets published in GBIF identified catalogue number formatting that was not handled in the clustering algorithm. This was presented to the group as a requirement to address. A fix to the issue identified was coded.
2. The clustered result occurrences were used to compare with the Plazi TreatmentBank datasets (publishingOrgKey = "7ce8aef0-9e92-11dc-8738-b8a03c50a862", as all datasets published by Plazi). Questions were first drafted on the relationship between digitised material citations from literature and their corresponding physical curationship. SQL queries were then issued to the databricks cluster in order to retrieve the answers in tabular form.
3. The clusters formed by the Meise Herbarium records published to GBIF were taken as a test case and further explored. The records published by Meise that clustered were also investigated on a taxonomic and spatial level.
4. Modifying the code to consider the possibility of multiple values being stored in the otherCatalogNumbers field. If multiple values were stored in this field, they should be split, then the individual component parts could be compared across other fields, potentially finding more clustering matches between records. A branch was made to explore this which functioned well on a smaller subsection of records, but encountered performance issues when run on larger quantities of data.

Results

The changes made during the week increased the count of records that link in GBIF from 43.7M to 50.5M with the ENA dataset increasing from 720k to 1.1M. Using the clustered result, as a practical application, we were able to acknowledge that Landcare Research, California Academy of Science and Museum national d'Histoire naturelle are the top three organisations that each holds more than 5000 specimens cited by literature digitised by Plazi. We were also able to further explore the comparison by grouping with the type status and georeferencing status which enables data quality check between collection metadata and material citations (https://bit.ly/gbif-clustering-plazi). We did not have time to fully analyse the impact and results of this beyond these simple metrics.

Conclusion

The team achieved the goal of a shared understanding of the current implementation operating at GBIF. The algorithm was improved with minor improvements but a mechanism to monitor improvements to the algorithm - or to assess new approaches - is still needed and alternative algorithms should be explored. A key outcome was the confirmation that using a shared cloud environment enables collaboration that otherwise would be difficult to achieve. This environment allows easy exploration of a large dataset, and having access to common, shared cloud computing capabilities with up-to-date exports of GBIF has great potential to enable easy exploration of GBIF data.

3.3.2. Registering biodiversity-related vocabulary as Wikidata lexemes and link their senses to Wikidata items (Topic 5)

Aim/problem/goal

A lexeme is the basic lexical unit of a language consisting of one or more words. Wikidata collects lexemes as structured data in any language. They allow for precise definitions and could potentially be used to extract meaning from texts during text mining. However, to populate Wikidata lexemes workflows are needed from text to Wikidata. This topic aims to create a workflow from TreatmentBank into Wikidata lexemes where they can then be enriched by the community.

Method

Two taxonomic treatments were selected as test input, one modern English taxonomic publication (Wongkamhaeng et al. 2020) and the other a 18th century German text on plant development (Goethe 1790). TextImager (Hemati et al. 2016) and TextAnnotator (Abrami et al. 2020) were used for Natural Language Processing (NLP) to extract biodiversity-related words and phrases, these were then uploaded as lexemes to Wikidata, there they were curated by adding information about their forms, senses and usages. We also visualised the lexemes in Ordia to assist with quality control, prioritisation and data exploration (Nielsen 2019).

Results

During the hackathon about thirty lexemes were created and annotated. An example is the word glabrous (https://www.wikidata.org/wiki/Lexeme:L593539) and the phrase deciduous forest (https://www.wikidata.org/wiki/Lexeme:L594039). These have then been annotated manually, for example by linking the lexemes 'deciduous' and 'forest' as the parts of the 'deciduous forest'.

Conclusion

Using existing natural language processing pipelines for part-of-speech tagging and semantic annotation of a given knowledge domain seems to be a viable approach to enable the automatic recognition and extraction of biodiversity-relevant terms and to convert them into Wikidata lexemes. If this is to be scaled up, further clarification is needed on which point in the workflow the community should be involved. Should users themselves be able to upload and process texts in the NLP pipeline? Should the service, including data output, be selectable for existing corpora (e.g. Biodiversity Heritage Library, Pensoft)? Which output format should be offered to keep the data transfer to Wikidata as simple as possible? Who offers and maintains the NLP service?

3.3.3. FAIR Digital Object design for data from multiple sources (Topic 6)

Aim/problem/goal

To design a workflow to create semantically enhanced FAIR Digital Objects that can interconnect disparate biodiversity data in the different research infrastructures within a coherent structure in the near future. Moreover, to provide a JSON-LD representation of an Open Digital Specimen (ODS) and an automated way to validate the structure of data against the ODS standard.

Method

An automated workflow was developed to create semantically enhanced Digital Objects validated against the new modelling framework created for ODS modelling. To this end, a Wikibase environment was set up and customised to develop the data types and properties for ODS. The Wikibases enabled the transformation of data types and properties from the DiSSCo's Modelling Framework into Shape Expressions (ShEx). A Digital Specimen is expected to satisfy this ShEx schema. Based on the ShEx schema, a workflow is designed to validate biodiversity data against the data model.

Results

- A framework and recipe to label schema and align terms.
- A new FAIR Digital Object schema and corresponding data types.
- A model/ontology based on a simple example of a Digital Specimen.
- A fully automated workflow from model to ShEx schema in Cordra and semantic validation of new digital specimen objects.

Conclusion

Even though JSON schema provides a structure, it cannot explicitly capture various aspects of the data such as type definition, constraints. RDF statements also can be incomplete or missing. It is important to explicitly articulate the schema (Kellou-Menouer et al. 2021). ShEx in this regard is more powerful than simple schema validation and can also help with the presentation of the semantics for both humans and machines. The use of Wikibase as a modelling framework has advantages for creating a future ODS standard compliant with Biodiversity Information Standards (TDWG) requirements due to better tracking of versioning.

3.3.4. Enriching Wikidata with information from OpenBiodiv about type specimens in context from different literature sources (Topic 7)

Aim/problem/goal

To develop a workflow that integrates knowledge about type materials from the OpenBiodiv knowledge graph with existing Wikidata records to enrich Wikidata with more data from biodiversity literature.

Method

The OpenBiodiv knowledge graph (Dimitrova et al. 2021) containing Linked Open Data statements extracted from literature through XML-tagging in publications was used as a data source. SPARQL queries were performed to OpenBiodiv to explore collections and institutions which have been used in the description of new taxa in the Biodiversity Data Journal (BDJ) using type specimens. Mapping of institutions and collections between OpenBiodiv, GBIF, Wikidata was done manually in some cases, when no identifier was available. Enrichment of Wikidata with information about type materials was done using OpenRefine.

Results

It was discovered that only about 2% (314 out of 14390) of all institutions and collections from the GBIF Registry of Scientific Collections (GRSciColl) are indexed in Wikidata with their GRSciColl identifier, with some being indexed without GRSciColl identifier. In addition, only 303 type specimen records were found in Wikidata. OpenBiodiv was used to discover information about type specimen locations and holding institutions/collections and map them to existing type specimen records on Wikidata, as well as create new ones, whilst keeping a reference to the original source (taxonomic article). Our contributions to Wikidata are accessible at: https://www.wikidata.org/wiki/Special:Contributions/ROMEnEwr.

Conclusion

The examination of existing records of institutions and collections in Wikidata and GRSciColl showed an ambiguous use of the term “collection”, as some institutions are regarded as collections and vice versa. A clear need to disambiguate institutions became evident, due to the multitude of identifiers used for a single institution in GRSciColl, as well as duplicate

institution code records. We suggest also a revision of Wikidata properties to help capture information about institutions and type specimens in a better way.

3.3.5. Linking specimen with material citation and vice versa (Topic 8)

Aim/problem/goal

Scholarly publications cite specimens on which the research is based. These, so-called, material citations, are in taxonomic literature, either within a taxonomic treatment, or in tables, often also including additional links such as genetic sequences with accession numbers. Traditionally these citations identify specimens in natural history institutions. Natural history institutions regularly publish their specimen details to GBIF together with their specimen identifiers. For us, linking directly to GBIF is an option that circumvents custom solutions for each institution, and at the same time allows the institutions to retrieve the links to the material citations via searching their uploaded occurrences and related, clustered occurrences – that is a material citation uploaded by TreatmentBank. We add the GBIF occurrence ID to the respective TreatmentBank record, and once concluded re-upload the dataset including the attributed material citation to GBIF.

Method

We developed an algorithm to link the material citations in the GBIF database and the specimens in the Natural History Museum of Bern (NMBE) collections. The algorithm hinges on calculating similarities (e.g. string to string edit distances for narratives, Euclidean distances for geolocalized data) between the instances in both sides of linking. Literature contents were harvested thanks to BICIKL services (i.e. Plazi, SIBiLS) and indexed to compare each material citation and specimen based on ex-ante selected attributes and calculates pairwise similarities accordingly. The attributes could be in the string type such as genus or family information, as well as numeric type such as latitude or longitude of the discovery place. The algorithm calculates the similarities for each data type separately and merges and normalises them at the end to find a final pairwise similarity in the interval of [0-1] between a material citation and a specimen. It sorts material citations for each specimen according to the similarity score. Finally, it assigns the most similar material citation's "material citation ID" to the corresponding specimen in the NMBE collection. Overall, the algorithm finds the most similar material citation for each specimen; thus, it bridges two datasets. Large curated sample data are now needed to fine-tune a data-driven matching method.

Results

We developed an algorithm for matching "material citations id" in GBIF to the NMBE specimen. A light graphic user interface has also been designed to support the manual discovery and curation of bi-directionnal links (Fig. 5).

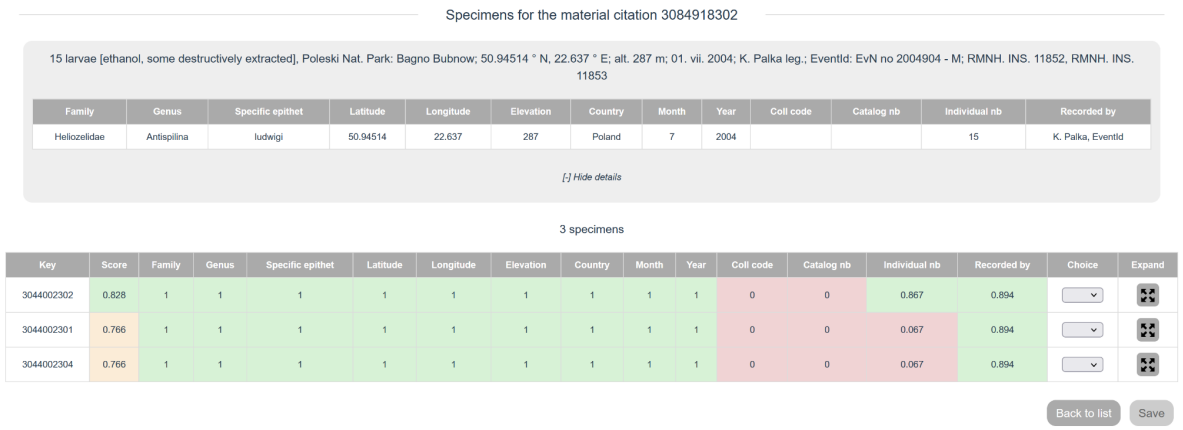

Figure 5. *Graphic user interface to support the cross-linking of citations and specimens.*

Conclusion

The use of GBIF as a broker of institutional databases has several advantages. First, only one bi-directional linking algorithm and interface has to be developed. The GBIF occurrences are continually, automatically updated, whenever an attribute has been added on the TreatmentBank side. The institutions are at ease when they want to update their records. The development of the clustering algorithm (see also Topic 3) will facilitate linking specimens and material citations from a particular institution in a first step – the searching over a billion occurrences will be time consuming and unsustainable for linking large numbers of material citations. This is even more complex due to the nature of material citations that can represent the entire specimen record to only parts, often in slightly different formats.

3.3.6. An IPFS-Blockchain Interface (Topic 10)

Aim/problem/goal

LifeBlock - based on Blockchain - technology will enable the management and organisation of data from various information sources with traceability, provenance, tokenization and application of FAIR principles. LifeBlock through its developing interfaces will allow the connection with infrastructures (e.g. GBIF, Plazi) participating in the BiCIKL project. The objective is to provide a unique platform to the community which addresses the provision of information from different information sources that can be enriched and extended in the future. The use of blockchain technology guarantees the performance of the afore-mentioned operations, as well as its strict compliance with FAIR principles. The inclusion of non-fungible tokenization (NFT) elements will allow information to be managed in different ways, including the generation of unique datasets that can be licensed and identified.

Method

Exchange of information using blockchain and InterPlanetary File System (IPFS) technologies with the infrastructures present at the hackathon: GBIF, Naturalis/DiSSCo, PlutoF, Plazi, OpenBioDiv.

Results

LifeBlock's data provenance and traceability system was refined by better understanding the characteristics of each of the data sources and identifying use cases of this technology in the themes of other hackathon teams. For DiSSCo (see also Topic 6), the LifeWatch ERIC team explored (1) The generation of traceability mechanisms on the events of changes in object information contained in its database. (2) The provision of traceability elements for images and storage space using the IPFS infrastructure is another identified user case. (3) Image comparison based on hashes for copy identification and to implement storage saving strategies. (4) Tokenization of digital objects to evaluate the possibility of introducing an NFT-based “micropayment” system for the environmental impact of the objects. For PlutoF the team identified potential use cases on the uploading, management and use of images and the traceability and storage of PlutoF curated data.

Conclusion

Compliance with FAIR principles and guaranteed mass storage of information are the main attributes that LifeBlock will be able to contribute to the community. Additionally, the encapsulation of information in a tokenized system based on NFT can bring numerous advantages to the management of the information generated. It will allow the inclusion of systems such as the “micropayments”, and the identification of the origin of the information and will allow the establishment of unique data sets whose generation and subsequent transmission can be permanently rewarded to the producers of information. Further exploration is ongoing.

4. Acknowledgements

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.

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.

5. References

Abrami, Giuseppe, Manuel Stoeckel, and Alexander Mehler. 2020. 'TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts'. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 891–900. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.112.

Briscoe, G. 2014. 'Digital Innovation: The Hackathon Phenomenon.' http://qmro.qmul.ac.uk/xmlui/handle/123456789/11418.

Clark, 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, Koenraad, Dimitris Koureas, and Peter Wittenburg. 2020. 'FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units'. Publications 8 (2): 21. https://doi.org/10.3390/publications8020021.

Dimitrova, Mariya, Viktor Senderov, Teodor Georgiev, Georgi Zhelezov, and Lyubomir Penev. 2021. 'Infrastructure and Population of the OpenBiodiv Biodiversity Knowledge Graph'. *Biodiversity Data Journal* 9 (September): e67671. https://doi.org/10.3897/BDJ.9.e67671.

Garcia, Leyla, Erick Antezana, Alexander Garcia, Evan Bolton, Rafael Jimenez, Pjotr Prins, Juan M. Banda, and Toshiaki Katayama. 2020. 'Ten Simple Rules to Run a Successful BioHackathon'. Edited by Scott Markel. *PLOS Computational Biology* 16 (5): e1007808. https://doi.org/10.1371/journal.pcbi.1007808.

Goethe, W. 1790. 'Der Versuch die Metamorphose der Pflanzen zu erklären.' Ettinger, Gotha. https://www.projekt-gutenberg.org/goethe/metamorp/metamorp.html.

Groom, Quentin, Anton Güntsch, Pieter Huybrechts, Nicole Kearney, Siobhan Leachman, Nicky Nicolson, Roderic D M Page, David P Shorthouse, Anne E Thessen, and Elspeth Haston. 2020. 'People Are Essential to Linking Biodiversity Data'. *Database* 2020 (November): baaa072. https://doi.org/10.1093/database/baaa072.

Groom, Quentin John, Mathias Dillen, Pieter Huybrechts, Rukaya Johaadien, Niki Kyriakopoulou, Francisco Jose Quevedo Fernandez, Maarten Trekels, and Wai Yee Wong. 2021. 'Connecting Molecular Sequences to Their Voucher Specimens'. Preprint. BioHackrXiv. https://doi.org/10.37044/osf.io/93qf4.

Güntsch, Anton, Roger Hyam, Gregor Hagedorn, Simon Chagnoux, Dominik Röpert, Ana Casino, Gabi Droege, et al. 2017. 'Actionable, Long-Term Stable and Semantic Web Compatible Identifiers for Access to Biological Collection Objects'. *Database* 2017 (January). https://doi.org/10.1093/database/bax003.

Guralnick, Robert, Tom Conlin, John Deck, Brian J. Stucky, and Nico Cellinese. 2014. 'The Trouble with Triplets in Biodiversity Informatics: A Data-Driven Case against Current

Identifier Practices'. Edited by Damon P. Little. *PLoS ONE* 9 (12): e114069. https://doi.org/10.1371/journal.pone.0114069.

Hemati, Wahed, Tolga Uslu, and Alexander Mehler. 2016. 'TextImager: A Distributed UIMA-Based System for NLP'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 59–63. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-2013.

Kalfatovic, Martin R., Grace Costantino, and Constance A. Rinaldo. 2019. 'The Biodiversity Heritage Library: Unveiling a World of Knowledge About Life on Earth'. In *Digital Libraries for Open Knowledge*, edited by Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt, 11799:352–55. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30760-8_32.

Kellou-Menouer, Kenza, Nikolaos Kardoulakis, Georgia Troullinou, Zoubida Kedad, Dimitris Plexousakis, and Haridimos Kondylakis. 2021. 'A Survey on Semantic Schema Discovery'. *The VLDB Journal*, November. https://doi.org/10.1007/s00778-021-00717-x.

Kõljalg, Urmas, Henrik R. Nilsson, Dmitry Schigel, Leho Tedersoo, Karl-Henrik Larsson, Tom W. May, Andy F. S. Taylor, et al. 2020. 'The Taxon Hypothesis Paradigm—On the Unambiguous Detection and Communication of Taxa'. *Microorganisms* 8 (12): 1910. https://doi.org/10.3390/microorganisms8121910.

Kollwitz, Christoph, and Barbara Dinter. 2019. 'What the Hack? – Towards a Taxonomy of Hackathons'. In *Business Process Management*, edited by Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling, 11675:354–69. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-26619-6_23.

McVeigh, Richard. 2022. 'Bacterial 16S Ribosomal RNA RefSeq Targeted Loci Project'. National Center for Biotechnology Information (NCBI). https://doi.org/10.15468/K2C8EN.

Medina Angarita, Maria Angelica, and Alexander Nolte. 2020. 'What Do We Know About Hackathon Outcomes and How to Support Them? – A Systematic Literature Review'. In *Collaboration Technologies and Social Computing*, edited by Alexander Nolte, Claudio Alvarez, Reiko Hishiyama, Irene-Angelica Chounta, María Jesús Rodríguez-Triana, and Tomoo Inoue, 12324:50–64. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58157-2_4.

Meeus, Sofie, Wouter Addink, Donat Agosti, Christos Arvanitidis, Mariya Dimitrova, Juan Miguel González-Aranda, Jörg Holetschek, et al. 2021a. 'Hacking Infrastructures

Together: Towards Better Interoperability of Infrastructures'. *Biodiversity Information Science and Standards* 5 (September): e74325. https://doi.org/10.3897/biss.5.74325.

Meeus, Sofie, Tom August, Maarten Trekels, Lien Reyserhove, and Quentin John Groom. 2021b. 'Network Analysis of Specimen Co-Collection'. Preprint. BioHackrXiv. https://doi.org/10.37044/osf.io/4ahng.

Nelson, Gil, Patrick Sweeney, and Edward Gilbert. 2018. 'Use of Globally Unique Identifiers (GUIDs) to Link Herbarium Specimen Records to Physical Specimens'. *Applications in Plant Sciences* 6 (2). https://doi.org/10.1002/aps3.1027.

Nielsen, Finn Årup. 2019. 'Ordia: A Web Application for Wikidata Lexemes'. In *The Semantic Web: ESWC 2019 Satellite Events*, edited by Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, et al., 11762:141–46. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-32327-1_28.

Page, Roderic. 2016. 'Towards a Biodiversity Knowledge Graph'. *Research Ideas and Outcomes* 2 (April): e8767. https://doi.org/10.3897/rio.2.e8767.

Penev, Lyubomir, Dimitrios Koureas, Quentin Groom, Jerry Lanfear, Donat Agosti, Ana Casino, Joe Miller, et al. 2021. 'Towards Interlinked FAIR Biodiversity Knowledge: The BiCIKL Perspective'. *Biodiversity Information Science and Standards* 5 (September): e74233. https://doi.org/10.3897/biss.5.74233.

Penev, Lyubomir, Dimitrios Koureas, Quentin Groom, Jerry Lanfear, Donat Agosti, Ana Casino, Joe Miller, et al. 2022. 'Biodiversity Community Integrated Knowledge Library (BiCIKL)'. *Research Ideas and Outcomes* 8 (January): e81136. https://doi.org/10.3897/rio.8.e81136.

Pe-Than, Ei Pa Pa, and James D. Herbsleb. 2019. 'Understanding Hackathons for Science: Collaboration, Affordances, and Outcomes'. In *Information in Contemporary Society*, edited by Natalie Greene Taylor, Caitlin Christian-Lamb, Michelle H. Martin, and Bonnie Nardi, 11420:27–37. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_3.

Richard, Gabriela T., Yasmin B. Kafai, Barrie Adleberg, and Orkan Telhan. 2015. 'StitchFest: Diversifying a College Hackathon to Broaden Participation and Perceptions in Computing'. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, 114–19. Kansas City Missouri USA: ACM. https://doi.org/10.1145/2676723.2677310.

Robbertse, Barbara. 2022. 'Fungal Internal Transcribed Spacer RNA (ITS) RefSeq Targeted Loci Project'. National Center for Biotechnology Information (NCBI). https://doi.org/10.15468/CM2GBP.

The International Barcode of Life Consortium (2016). International Barcode of Life project (iBOL) Barcode Index Numbers (BINs). Checklist dataset https://doi.org/10.15468/wvfqoi accessed via GBIF.org on 2022-02-15.

Vrandečić, Denny. 2012. 'Wikidata: A New Platform for Collaborative Data Collection'. In *Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion*, 1063. Lyon, France: ACM Press. https://doi.org/10.1145/2187980.2188242.

Wikipedia Weekly. 2021. *Wikipedia Weekly Network - Live Editing Wikipedia: Biodiversity Edition #6*. https://youtu.be/wnHs1pLyoPU.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wissemann, Volker. 2007. 'Plant Evolution by Means of Hybridization'. *Systematics and Biodiversity* 5 (3): 243–53. https://doi.org/10.1017/S1477200007002381.

Wongkamhaeng, Koraon, Pongrat Dumrongrojwattana, Myung-Hwa Shin, and Chaichat Boonyanusith. 2020. 'Grandidierella Gilesi Chilton, 1921 (Amphipoda, Aoridae), First Encounter of Non-Indigenous Amphipod in the Lam Ta Khong River, Nakhon Ratchasima Province, North-Eastern Thailand'. *Biodiversity Data Journal* 8 (March): e46452. https://doi.org/10.3897/BDJ.8.e46452.

6. Appendix

Figure 6. *On-site participants of the BiCIKL hackathon at Meise Botanic Garden.*

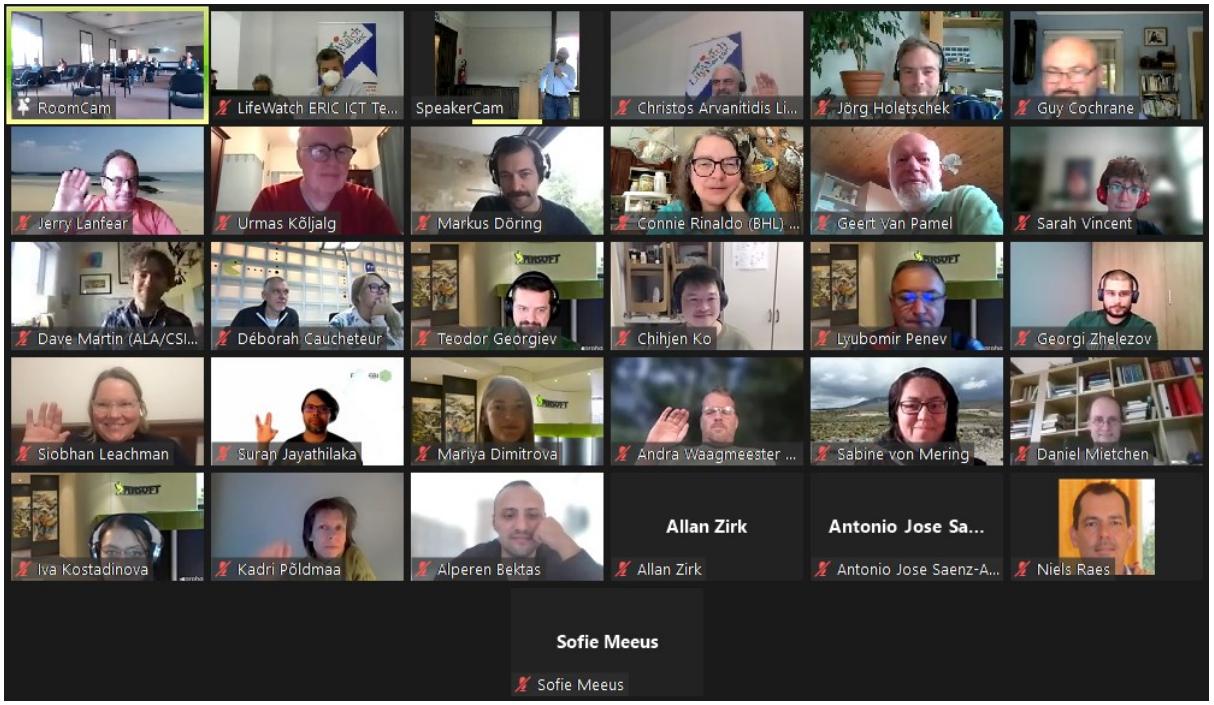

Figure 7. *Online participants of the BiCIKL hackathon at Meise Botanic Garden.*

Table 1: *List of topics tackled during the BiCIKL hackathon and the BioHackathon Europe with links to the associated GitHub repositories.*

Topic number	Topic title	Hackathon	GitHub repository
1	Finding the lost parents	BiCIKL	link
2	How good are Triple IDs in ENA?	BiCIKL	link
3	Enhance the GBIF clustering algorithms	BiCIKL	link
4	Assigning latin scientific names to OTUs based on sequence clusters	BiCIKL	link
5	Registering biodiversity-related vocabulary as Wikidata lexemes and link their senses to Wikidata items	BiCIKL	link
6	FAIR Digital Object design from multiple sources	BiCIKL	link
7	Enriching Wikidata with information from OpenBiodiv about type specimens in context from different literature sources	BiCIKL	link
8	Linking specimen with material citation and vice versa	BiCIKL	link
9	Hidden women in science	BiCIKL	link
10	An IPFS-Blockchain Interface	BiCIKL	link
11	Join the dots: Making sense out of biodiversity data with a human focus	BioHackathon Europe	link
12	CAB2: A step towards Biodiversity data enrichment	BioHackathon Europe	/

Table 2: *List of participants involved in the BiCIKL hackathon and the BioHackathon Europe.*

First name	Last name	Affiliation
Sofie	Meeus	Meise Botanic Garden
Quentin	Groom	Meise Botanic Garden
Pieter	Huybrechts	Meise Botanic Garden
Niki	Kyriakopoulou	Naturalis Biodiversity Center
Patricia	Mergen	Meise Botanic Garden
Kat	Thornton	Science Stories
Sabine	von Mering	Museum für Naturkunde Berlin, Germany
Siobhan	Leachman	Aotearoa New Zealand Wikimedia user group
Mathias	Dillen	Meise Botanic Garden
Patrick	Ruch	SIB & HES-SO
Maarten	Trekels	Meise Botanic Garden
Jörg	Holetschek	BGBM
Pierre-André	Michel	SIB
Sharif	Islam	Naturalis Biodiversity Center
Guido	Sautter	Plazi
Jonas	Grieb	Senckenberg (Frankfurt, Germany)
Mariya	Dimitrova	Pensoft
Lyubomir	Penev	Pensoft
Georgi	Zhelezov	Pensoft
Constance	Rinaldo	Biodiversity Heritage Library
Teodor	Georgiev	Pensoft Publishers
Wouter	Addink	Naturalis
Déborah	Caucheteur	SIB
Dave	Martin	CSIRO / ALA
Claus	Weiland	SGN, DiSSCo technical team
Marina	Golivets	Helmholtz Centre for Environmental Research - UFZ
Ben	Scott	Natural History Museum
Sarah	Vincent	Natural History Museum, London
Guy	Cochrane	
Christine	Driller	Senckenberg – Leibniz Institution for Biodiversity and Earth System Research
Tim	Robertson	GBIF
Thomas	Jeppesen	GBIF / COL
Nicky	Nicolson	RBG Kew
Tobias	Frøslev	GLOBE Institute, University of Copenhagen / UNITE
Donat	Agosti	Plazi
Rukaya	Johaadien	GBIF Norway
Visotheary	Ung	MNHN - TDWG
Christos	Arvanitidis	LifeWatch ERIC
Jerry	Lanfear	ELIXIR Hub
Kessy	Abarenkov	University of Tartu Natural History Museum

Allan	Zirk	UTARTU
Urmas	Kõljalg	Professor
Boris	Barov	Pensoft
Alexander	Wolodkin	Senckenberg – Leibniz Institution for Biodiversity and Earth System Research
Beat	Estermann	Bern University of Applied Sciences
Suran	Jayathilaka	ENA
Alperen	Bektas	Bern University of Applied Sciences
Vishnukumar	Balavenkataraman Kadhirvelu	ENA
Andra	Waagmeester	Micelio / GeneWiki
Geert	Van Pamel	Wikimedia België
Markus	Döring	GBIF / COL
David	Fichtmueller	Botanic Garden and Botanial Museum Berlin (BGBM)
Francisco Manuel	Sanchez Cano	LifeWatch ERIC
Antonio Jose	Saenz-Albanes	LifeWatch ERIC
Joaquin	Lopez Lerida	Lifewatch ERIC
Pablo	Guerrero	LifeWatch ERIC
Juan Miguel	González Aranda	LifeWatch ERIC
Chihjen	Ko	Independent consultant
Tom	August	UK Centre for Ecology and Hydrology
Henry	Engledow	Meise Botanic Garden
Lien	Reyserhove	Research Institute for Nature and Forest (Belgium)
Marcus	Guidoti	Plazi
Giulia	Agostinetto	University of Milan
Daniel	Mietchen	RIO Journal
Simon	Rolph	UK Centre for Ecology and Hydrology
Alberto	Brusati	University of Milan
Anna	Sandionigi	Quantia Consulting
Dario	Pescini	University of Milan
Kenzo	Milleville	Ghent University
Krishna	Chandrasekar	Ghent University
Kadri	Põldmaa	University of Tartu
Niels	Raes	Naturalis Biodiversity Center