SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

ADVANCES IN BIODIVERSITY INFORMATION STANDARDS AND PROCESSES (DELIVERABLE 4.1)

MAARTEN TREKELS, ANTON GÜNTSCH, ROGER HYAM, MANUEL SÁNCHEZ, SABINE VON MERING, CLAUS WEILAND, MATT WOODBURN AND QUENTIN GROOM

Grant Agreement Number | 823827
Acronym | SYNTHESYS PLUS
Call | H2020-INFRAIA-2018-2020
Start date | 01/02/2019
Duration | 48 months
Work Package | 4
Work Package Lead | Quentin Groom
Delivery date | 14.10.2022

Contents

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+
Synthesis of Systematic Resources
a DiSSCo project

Introduction

Beginning at the turn of the millennium biodiversity informatics has emerged as a major focus for museums, herbaria and other biological collections (Bisby 2000). This has been fuelled by the excellerating biodiversity crisis and the consequent need to monitor, predict and understand the biodiversity of the planet. Furthermore, the 1993 Convention on Biological Diversity and the subsequent Aichi Biodiversity Targets and Nagoya Protocol have necessitated a digital approach to biodiversity conservation to be able to respond to the demand for more reliable information at ever greater speed. Since 2004, SYNTHESYS has been at the forefront of the digital transformation of biodiversity collections and in its latest iteration this has been at the core of all work on the project (Smith et al. 2019). Networking action 4 on digital standards and processes has links to all other aspects of the project, including the training, access to collections, molecular standards and processes, and links to other communities and internationalisation.

Other initiatives that have paralleled the digital transformation of collections, including the push towards increased reproducibility of research (Baker 2016); the desire to preserve and reuse the data from research projects (Borgerud and Borglund 2020); the need to make digital data more findable, accessible, interoperable and reusable, as encapsulated by the FAIR Data Principles (Wilkinson et al. 2016), and the needed use of Linked Open Data to connect all of the entities of biodiversity science (Penev et al. 2019).

For their part, biodiversity collections have been imaging their specimens in great volumes and making these images and the associated metadata available openly online. Thus, creating an enormous volume of data on biodiversity that we have the challenge to make available and analyse. SYNTHESYS+ has stepped up to these challenges by supporting standards and process development on a number of fronts, engaging collections of all sizes in the development and adoption of standards.

The choice of standards on which we have worked on has been very much a bottom-up process based upon the needs of collections and the taxonomy community. For example, the need to compare specimens from multiple collections online has been driving the need for IIIF technology and the need to report on and document the process of digitization has been the driving force behind MIDS and Latimer Core.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

In this report we document the different work streams that Networking Action 4 has been supporting under SYNTHESYS+. We outline the problems we are trying to solve, the way that we are trying to solve them, the progress we have made, and finally we make recommendations for how collections can progress their biodiversity informatics agendas.

Increased findability of collections

Initiatives to assess the data on natural history collections will highly benefit from a data standard on collections information. In order to discover the huge amount of information on worldwide biodiversity which is present in our natural science collections, it is essential to be able to compare and aggregate data on our collections. The use-cases for a (meta-)data standard on collection information range from the very high-level information on the collection as a whole, down to the level of (small) groups of specimens inside the subcollections. The level at which this information can be provided is very dependent on the digitization level of these, but the data standard should make it possible to gather basic information on all collections. This will increase the visibility of collections significantly, but also allows users to track down the information they need. Examples of these users are data aggregators, scientists, policy makers…

The SYNTHESYS+ project actively supported the development of the new data standard which currently is named 'Latimer Core' (Woodburn 2022). Through the TDWG Collections Description Interest Group, the structure of the data standard was proposed and terms were defined. The project facilitated several in-person and online meetings to advance the development of the data standard. The strategy was chosen to organise weekly 'virtual barbecues', that were free for people to join according to their availability. This ensured that the progress on the development of the standard was continuous throughout the whole process. In order to collect all feedback of the stakeholders of the standard, all proposed terms and definitions in the standard were tracked with a GitHub repository (https://github.com/tdwg/cd).

Parallel to the development of the standard itself, the use of a Wikibase instance was piloted as a means to provide an exemplar implementation (see next section). By providing an implementation of the standard that can be used by non-technical people, it was possible to highlight potential shortcomings of the proposed standard early on in the development.

At the time of writing this report, the first version of the Latimer Core standard was submitted to TDWG for formal review. A schematic representation of the different proposed classes is shown in figure 1. Using the Wikibase instance showed a great potential for usage in future standards development. Therefore, including an instance of the Wikibase in the documentation of the standard is foreseen.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

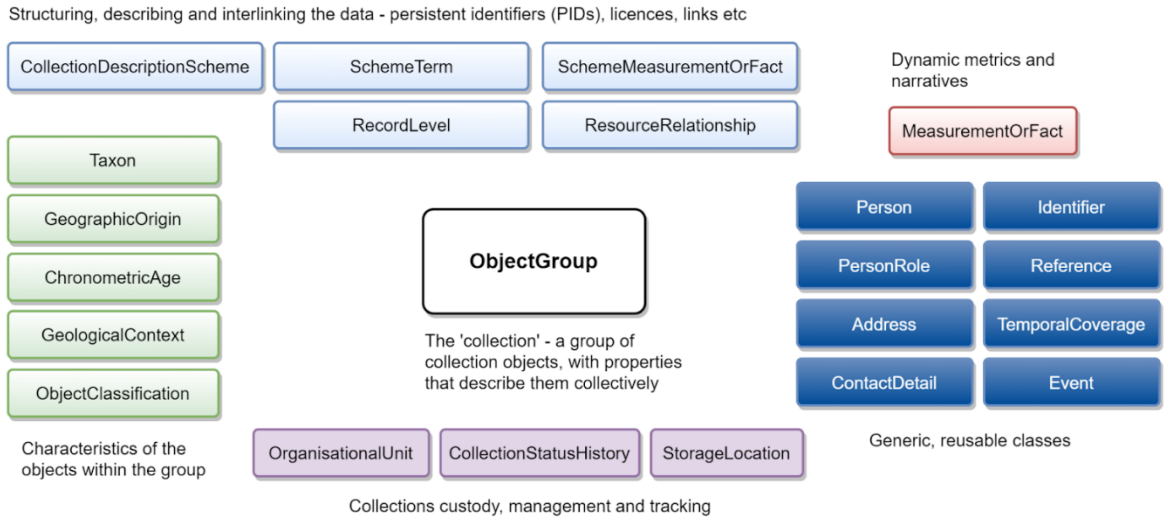


Figure 1: Structure of the Latimer Core classes (from (Woodburn 2022)).

Facilitating standards development through Wikibase

The process of developing a data standard is complicated by the large number of stakeholders, the legacy of existing data and difficulty of testing the proposed new standard. What frequently looks good in draft, fails when confronted with real data and actual users. What is needed is a simple platform that is readily available to human and machine users to be able to read and write to where a standard can be tested, reconfigured and updated.

Over the course of the project, Wikibase was used in two different contexts. Firstly, it was used during the EU project DiSSCO Prepare to build a Wikibase-based modelling framework. This has the possibilities of semantic descriptions and linking but at the same time can also be used by people with little ontology experience (Fichtmueller & Güntsch 2021, 2022). Wikibase is an extension to Mediawiki which is the software running both Wikipedia and Wikidata. With Wikibase, knowledge bases can be flexibly developed that go beyond the original context of Wikidata and can be operated in separate instances.The "DiSSCo modelling Framework" is such an application that is used for data modelling processes and the semantic backbone for processing specimen data in the DiSSCo context. The development of the platform is based, among other things, on requirements that are important in the SYNTHESYS+ project, e.g. in the development of MIDS and the CETAF Specimen Preview Profile (CSPP).

Secondly, the potential of using a Wikibase was tested as an easy platform to engage non-technical users in the development of a data standard. During the development of the Latimer Core standard, a Wikibase instance was set up to be used as a sandbox containing the newly developed term within the standard (https://tdwg-cd.wikibase.cloud/). This allowed potential users of the data standard to test different use-cases with an initial implementation of the data standard (Grant et al. 2020, Trekels et al. 2020). The graph nature of the Wikibase implementation allowed for sufficient

flexibility to test different implementation strategies. Within the TDWG task group it is decided that an instance of Wikibase can be included in the non-normative documentation of the standard (figure 2). When updates to the standard are proposed, it can be tested in the Wikibase.

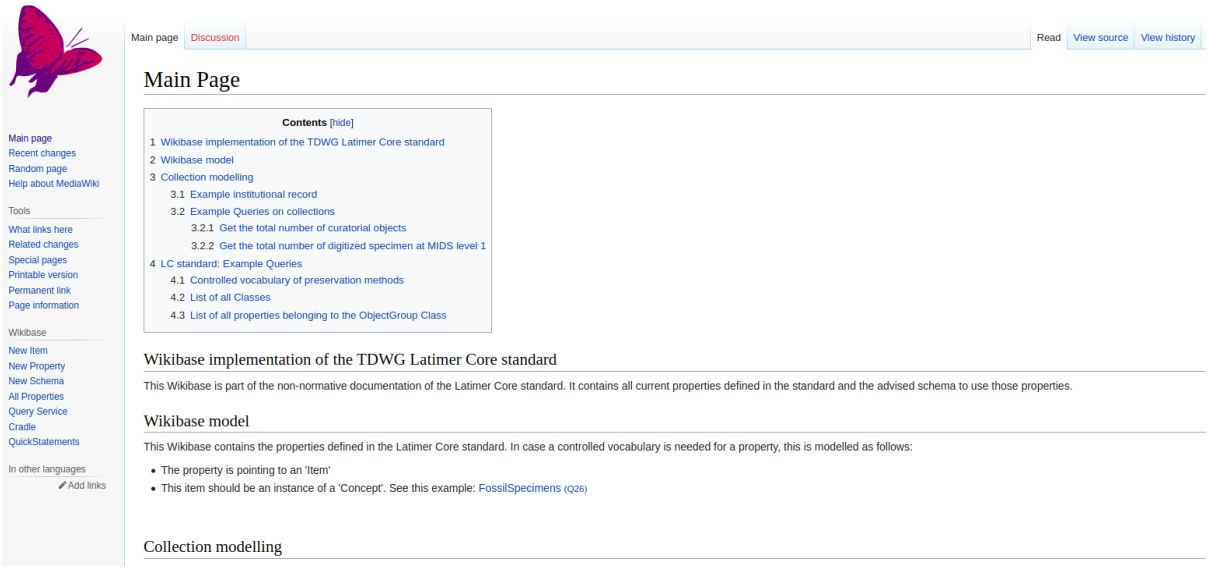


Figure 2: Screenshot of the Wikibase implementation of Latimer Core.

The DiSSCo modelling Framework is fully functional and in active use in semantic modelling for the DiSSCo infrastructure. Use of the platform for other modelling initiatives, especially in the context of TDWG Biodiversity Information Standards, is envisaged. With the development of the Latimer Core standard, it is planned to make a Wikibase instance part of the documentation upon releasing the first version of the standard.

Digital specimens

The digitization of the world's biodiversity collections will take many years and even when the backlog of specimens is digital, all new specimens need to be digitised before they can be shelved. Monitoring and reporting on the state of digitization is therefore a priority as it will determine priorities for collections and countries. Currently, there is no way to express the level of digitization of a specimen, nor compare the levels of digitization between collections.

Additionally, large scale digitization efforts have the objective to represent the physical specimen with comprehensive associated material in the digital domain, including two-/three-dimensional digital images of the specimen plus accompanying images like computerised tomography scans, measurements of associated materials like parasites, microbiomes and environmental samples, labels and markings on the physical specimen itself converted to machine-readable text or digital recordings of discovery sites (Nelson 2018). Major challenges arising from these efforts are the development of interoperable data standards and models as well as building up the information infrastructures for exchange and mobilisation of collection data between different research fields.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

The Minimum Information about a Digital Specimen (MIDS) standard is intended to fill this gap in knowledge. The draft MIDS standard defines four levels of digitization from the very basic level of digitization where only the accession number and the institution are digital, to a fully digitised specimen where both an image is available, and core metadata on where, when and who collected the specimen. Armed with this information, collections can be compared and digitally unified with each other.

Correspondingly, DiSSCo developed the Digital Specimen data model in close alignment with the MIDS specification to realize the FAIR-compliant integration of digitized collection data into hyper infrastructures like the European Open Science Cloud (Wilkinson 2016, Islam 2020). Digital Specimens are typed compound objects of Persistent Identifier, Metadata and content embedded within the wider cross-domain framework of FAIR Digital Objects (De Smedt 2020, Wittenburg 2022).

A Digital Specimen encapsulates and persistently links to relevant information artefacts (i.s. gathered during digitization), which are about the physical specimen like sequences stored in INDSC databases, images in DiSSCo's Digital Specimen repository and occurrence data in GBIF. The object-centred representation of the logical structure of a specimen's data enables high-level operations producing more derived data about the specimen, e.g. calculation of the MIDS level or (Deep Learning-based) feature tracking and extraction from accompanying image objects (Hardisty 2022, Grieb 2021).

To design and implement Digital Specimens and other biodiversity FDO types (e.g. covering image capture) a technical specification was initiated within the DiSSCo Prepare Project: the Open Digital Specimen specification (openDS). Further development of openDS is highly aligned with key specifications and infrastructures in SYNTHESYS+:

- MIDS's information elements to describe specimens within a digital framework are mapped to attributes of the core data types provided by openDS (e.g. ods:DigitalSpecimen). Higher MIDS levels (2+) are intended to use the openDS ontology extensively.
- the Specimen Data Refinery (SDR), a cloud-based platform for processing specimens involving Machine-Learning pipelines. OpenDS provides the data models and types to capture derived information like extracted trait data (e.g. subclasses of ods:MediaObject).

A regular coordination meeting ("openDS breakout group") was established with key developers and modellers of the DiSSCo-linked projects in concern to organise the continuous assessment of requirements for openDS . Relevant development progress is published in the collaborative DiSSCo Modelling Framework (Fichtmueller & Güntsch 2021, 2022). Currently openDS provides the mappings up to MIDS 1.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

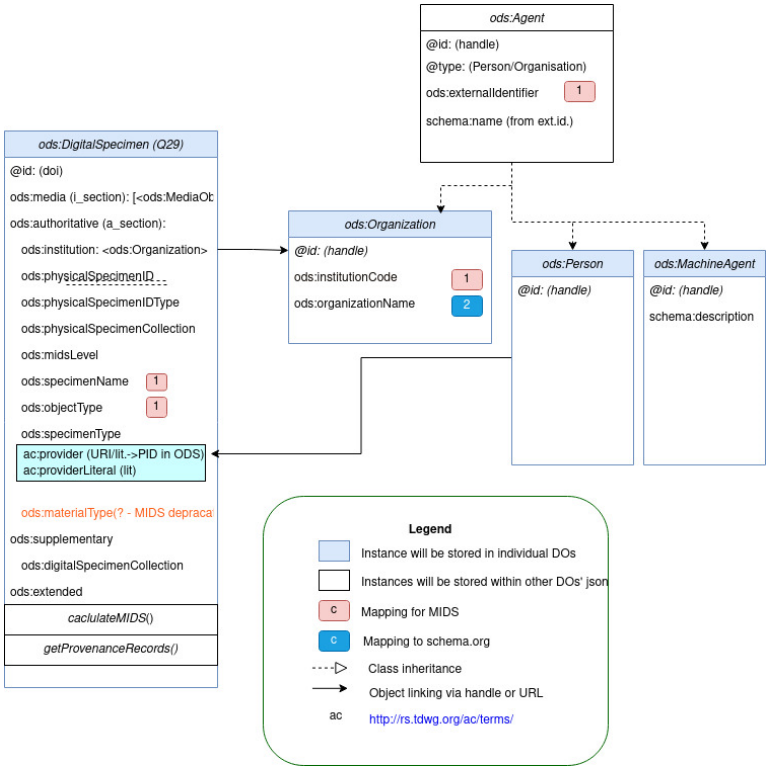


Figure 3: Outline (excerpt) of core classes of openDS including ods:DigitalSpecimen and ods:Agent, a parent class for organisation, person and machine agents. Mappings to MIDS and schema.org are shown here (pink/blue boxes). Actual development of the data models takes place in the DiSSCO modelling Framework (https://modelling.dissco.tech).

Leveraging data enrichment

Specimens have many connections to other kinds of data, including to other specimens. This helps them provide their function as vouchers for research, as well as a means to support new research. For example, nomenclatural type specimens are connected to taxonomic names, and these and other specimens are cited in taxonomic and other literature. Specimens are also linked to the places and dates where they were collected, and are linked to their physical traits (e.g. female) and to characteristics of the taxon (e.g. dioecious). However, one of the most common and consistent links is to the people who collected and determined their identity.

Traditionally, specimens, such as type specimens, were distinguished in literature by their collector name, the collector's number, the collection where they were deposited and sometimes the date or year of collection. Nevertheless, these details do not necessarily ensure the specimen referred to is unique and there is also considerable variability and ambiguity in the way collector's names, collections, collector numbers and dates are formatted. In order to create persistent, bidirectional links between people and specimens we need long lasting identifiers for both the specimens and the people.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Needless to say, creating persistent, globally unique, identifiers for specimens has been a goal of biodiversity informatics for some time (Clark, Martin & Liefeld 2004). Solutions requiring a unique UUID (Leach et al. 2005) or a resolving service (Klump & Huber 2017) have failed to gain traction in the community, however, at least among European collections the so-called "CETAF Stable identifier" has proved popular. This identifier emerged as a suggestion of (Hyam, Drinkwater & Harris 2012) to use URIs to permanently identify a specimen and was further developed by (Güntsch et al. 2017). The concept is to use URIs to indicate a landing page for specimen details on the internet. The URI is a redirect to the actual location of information and can be kept stable even if the specimen information is moved. The URIs stability is underwritten by the institution who is its guardian and relies on the DNS system of the internet to redirect URIs and on the institutional willingness to maintain its stability.

In the case of people, there are a wide variety of identifiers in use. For example, the Virtual International Authority File (VIAF) is created by a consortium of libraries to identify authors, mainly of books. However, there are many other bibliographic authority files, with different focuses. Some are global, some are national and others are institutional. In the case of living research scientists we have the ORCID identifier system. ORCID profiles are managed by the person it refers to, but can contain information of the persons institutional affiliations, grants awarded and their publications. ORCID is an open, reliable system, with a clear sustainability plan and no internal restrictions to its use globally. It is therefore a good solution to identifying living people uniquely. Nevertheless, many of the people associated with collections are dead, which rules out the use of ORCID, except for those people who had registered for an ORCID Identifier before they died.

Although many collectors and identifiers of specimens were also authors, by no means were all of them, which means authority files, such as VIAF are not a universal solution. Alternatively, there are databases of botanists (e.g. HUH) and entomologists (e.g. Entomologists of the World), which also provide identifiers. These are quite comprehensive, but are only suitable for some taxonomic groups and furthermore, cannot be added to or amended if people are missing or details are wrong. Therefore, a universal solution is needed that combines the numerous sources of people data and identifiers into one source. This is where Wikidata fills a need. It contains identifiers from numerous authority files and other biographical databases, but also contains referenced biographical data for people. The notability requirements for Wikidata are low enough so that the majority of people are notable if they have been mentioned in a publication or specimens collected by them are housed in a scientific collection.

Given access to stable identifiers both for people and specimens we are able to bidirectional link these entities and examine what additional information emerges from their union.

During SYNTHESYS+ partner institutions either implemented CETAF Stable Identifiers for their collection specimens, or improved their degree of compliance with the standard. This was achieved through technical support of partner institutions and through improving the documentation (https://cetafidentifiers.biowikifarm.net/wiki/Main_Page). The current compliance level of NA4

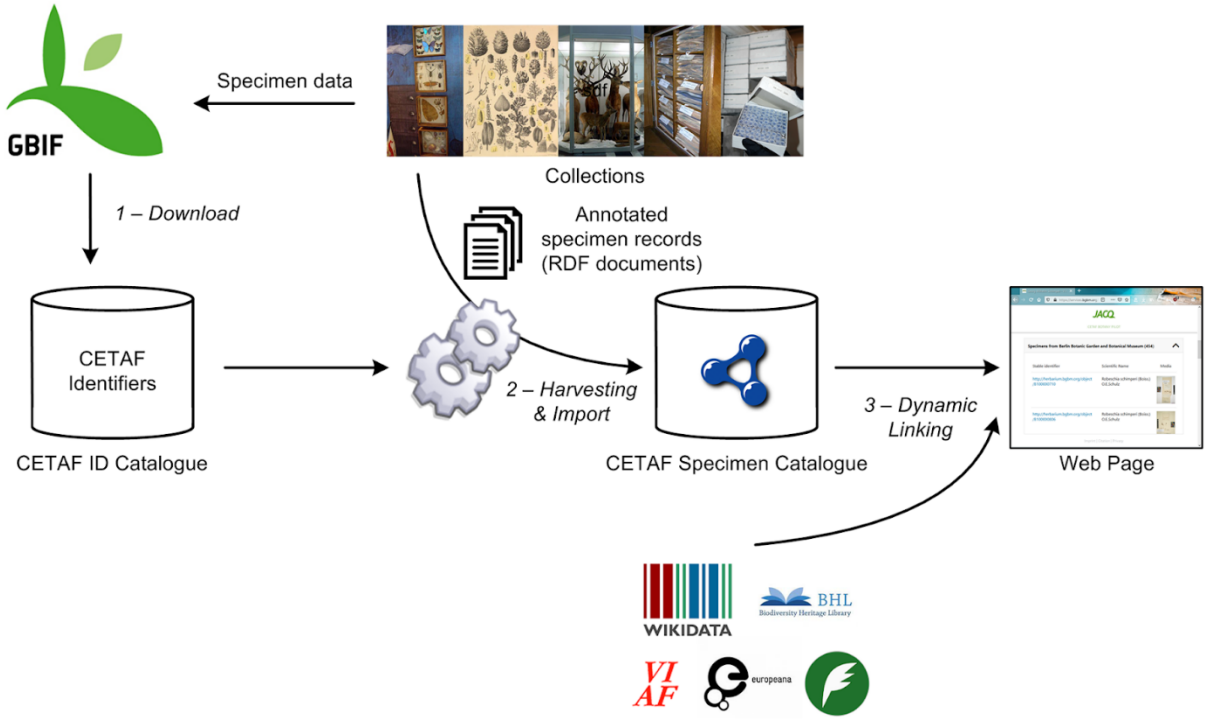
SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

partners can be seen on the standards_compliance_dashboard (https://cetafidentifiers.biowikifarm.net/wiki/Standards_compliance_dashboard).

Some NA4 partner institutions also went through their most prolific collectors in their collection management systems and added a stable identifier for those people, ultimately publishing those identifiers into GBIF using the recordedById field of Darwin Core. Once published the data could be harvested in bulk from the partner institutions and merged into a RDF triple store (Güntsch et al. 2021) (figure 4). From this triple store it is possible to create comprehensive collations of a person's collections from multiple institutions, but also, using the same identifiers, combine their specimen data with biographical details from Wikidata and elsewhere.

During SYNTHESYS+ 51 collections were brought into the CETAF Stable Identifier system, delivering an additional 50 million+ specimens that can be referred to uniquely. Furthermore, the system and the collections adopting it have been thoroughly documented (https://cetafidentifiers.biowikifarm.net/). We have published four papers on the use of person identifiers in collections and run several workshops, particularly in conjunction with the COST Mobilise Action (Groom et al. 2019, 2020 & 2022, Güntsch et al. 2021). We have also been actively encouraging the use of the people disambiguation system Bionomia (https://bionomia.net/), which allows people to claim their own specimens and help others. Going beyond SYNTHESYS+ we see this work only expanding to new collections and for the results to start spawning new applications and research.



SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Figure 4. A schematic of the workflow from collections and GBIF to the display website with merged specimen, bibliographical and biographic data for a person. GBIF provides a list of the currently available CETAF Specimen IDs (step 1), which are then used to harvest and import the corresponding specimen data of the collections into an RDF triple store (step 2). This provides the anchor point for generating dynamic web pages for people (step 3). From (Güntsch et al. 2021).

Data standards and societal changes

There is an enormous quantity of biodiversity knowledge in printed literature, but access to that literature is challenging when it is not digital. This is particularly true for people far from dedicated taxonomic libraries, but even more during a pandemic lockdown when no one had access to such libraries. Furthermore, that literature has many implicit links to specimens, places, taxa and people. If those links could be made explicit and digital, they would enormously increase the findability of data, and create a whole new source of research data.

The COVID-19 pandemic had a profound impact on everyone's lives, but it was also a challenge to scientists to find better ways to share data related to zoonosis. At the beginning of the pandemic, as countries across the world went into lockdown, the Consortium of European Taxonomic Facilities (CETAF) and the Distributed System of Scientific Collections (DiSSCo) joined forces to set up a COVID-19 Task Force that many SYNTHESYS+ partners contributed to. We particularly worked on the themes of “construction of a knowledge base relevant for pandemics” and “metadata registering practices”. These teams were exceptionally diverse collaborations and included participants from many disciplines and countries. They met online every week and set their own agendas for what topics they saw as most important.

The focus of the knowledge base team soon centred around the availability of biotic interaction data for bats. This was seen as a significant knowledge gap that prevents us understanding the evolution and spread of disease in wild populations. It led to indexing of some large datasets on the interactions of bats, viruses and other species on the Global Biotic Interactions (GloBI) database (e.g. Groom & Poelen 2020).

The team came to two important conclusions both from our discussions and experimentation with the data. Firstly, there is a vast wealth of scientific data that is effectively locked away in scientific publications and unavailable for further analysis. This is particularly true for data on biotic interactions. We therefore advocate in (Upham et al. 2021) for the creation of data extraction pipelines to extract these data en masse and the connection of these data together using persistent identifiers. In this paper we also promoted improvement in the way that newly published research is linked to external entities such as taxonomy, nomenclature, people, places and other literature.

Secondly, biodiversity science is often kept separate from disciplines related to human activities, such as farming, forestry and fisheries. Whereas we argue that this is a false dichotomy. Cultivated plants, domestic animals and captive animals are all part of the ecosystem. Without a holistic view of ecosystems, that includes all organisms you can not come to conclusions about how biodiversity

will change in the future and what impacts global environmental change will have on humans (Groom et al. 2021). Biodiversity informatics needs to learn from the 'One Health' approach that human wellbeing and biodiversity are interwoven and should be considered as a whole, rather than separate entities (Atlas 2012).

Curating big datasets by bridging between communities

Image corpora

Owing to large digitization projects running in many natural history collections, a vast number of digital images have become available to the community. Other communities that are dealing with large corpora of images include the archives and library community. This community developed the IIIF standard. IIIF is a set of API specifications that are used to display high resolution images. It is an established technology but little used in the natural history community. If it was more widely adopted it would facilitate the construction of shared browse, analysis and annotation tools. Another benefit of using established data standards is sustainability. By increasing the community of users involved in the development of the standard, the chance of it becoming obsolete in the short term is negligible.

In order to bridge the gap, several exemplar implementations were developed in recent years (Hyam 2019, 2021). The potential of this approach is immense. Virtual catalogues can be built using the images of several participating institutions. Moreover, the specimen images can be used more easily in other contexts such as the digital humanities. By incorporating the IIIF manifests in the large infrastructures (such as data aggregators like GBIF or Europeana), images become more visible and accessible to a wider public.

During the course of the SYNTHESYS+ project, 10 exemplar IIIF implementations have been created by the participating collections (Hyam 2021). In order to demonstrate the possibilities of IIIF in natural history collections, a pilot project specifically focussing on herbaria was developed. This project shows the capabilities of annotating and comparing specimens in a unified manner between collections. Another achievement is related to the inclusion of the IIIF images in GBIF. The occurrence records of the specimens on GBIF can now also include a link to the IIIF resource, and it becomes much easier to find and compare specimens to each other.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

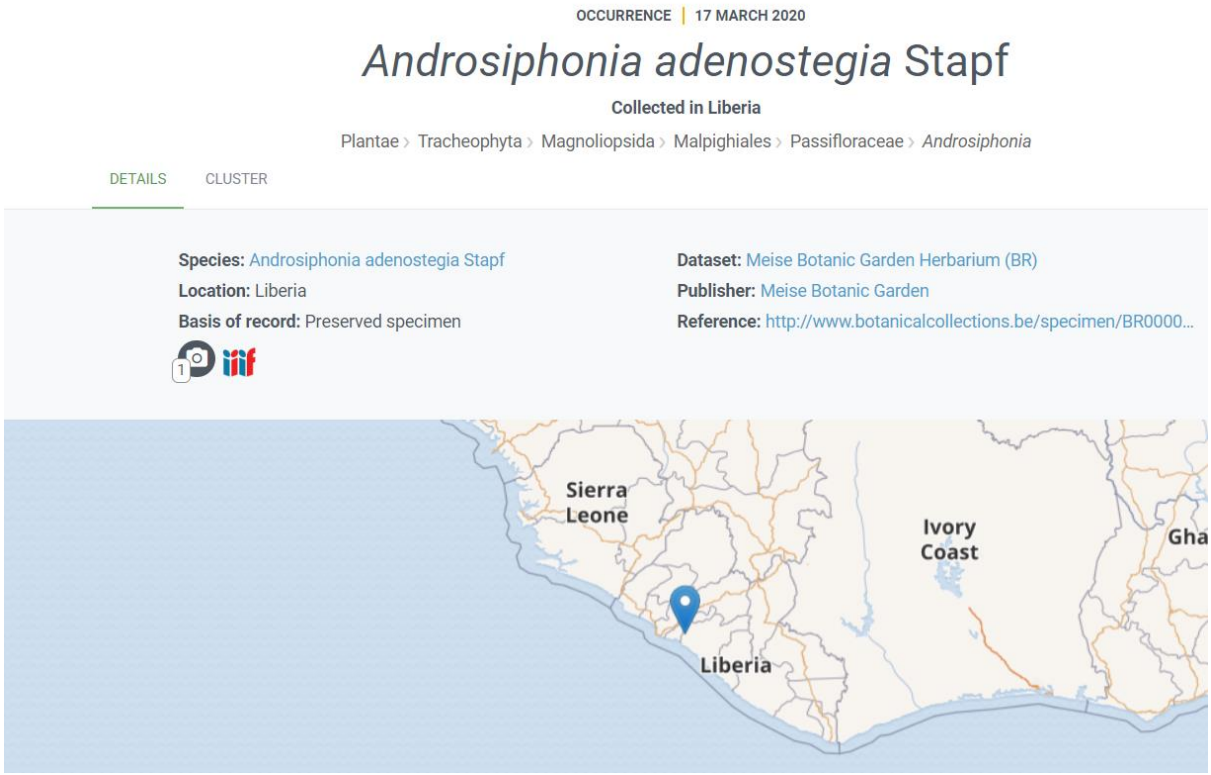


Figure 5: Inclusion of IIIF manifests in the GBIF dataset of Meise Botanic Garden.

Data curation

An additional effect of the ongoing mass digitization of specimens, is the vast amount of data that becomes available for research. Managing, curating and cleaning of these data is a time consuming process, which is often beyond the normal workload of curators, database managers or scientists. Moreover, a lot of the data elements we use are common across institutions, particularly publications, geography and people. It is a large waste of resources if data on these entities have to be maintained locally.

Engaging with large volunteer networks is also an essential step to make in the biodiversity standards community. Knowledge on biodiversity is more and more collected and shared through platforms such as the ones provided by the Wikimedia Foundation. Wikidata is a structured data source that contains about 100 million data items about the world, including information about people, collection holding institutes, taxa, etc. By engaging with this community, it is possible to align commonly used standards in biodiversity research to data available through Wikidata. This will facilitate the data flow into repositories such as WIkidata and allow an active enrichment of the current data that is held by the institutions.

Engaging volunteers needs not only efficient and simple platforms that facilitate this work, but also defined workflows and processes to deal with these enrichments. During SYNTHESYS+, the work

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

was started to define the workflows, tools and processes dealing with the disambiguation of people (collectors) inside the collection data. Three platforms were identified as crucial in the disambiguation of people: Wikidata, ORCiD, and Bionomia.

Future steps for the biodiversity standards community

Building on the lessons learned from the SYNTHESYS+ project, we are proposing 4 objectives for the biodiversity standards community. These are key areas where we can improve the relevance of natural history collections, not only in the domain of biodiversity information science, but also in relation to other disciplines. As such they can serve a wide variety of purposes within science, industry, policy…

Objective 1: To ensure that specimens in any collection can be referenced uniquely

A lot of the information (taxonomically, geographic distribution, collector…) on biodiversity is attached and originating from the specimens. To ensure that this base data is preserved and linked to the science outputs, it is essential to be able to reference uniquely to the (digital) specimens. On the one hand to the physical specimen, on the other hand the digital representation of that specimen. We propose therefore to roll out the system of CETAF stable identifiers to (almost) all CETAF institutions and beyond. In case smaller institutions don't have the infrastructure available, the CETAF network should look into sharing resources to enable the implementation of the identifiers.

Regarding the digital specimens, the TDWG community should engage in the development of an "extended digital specimen" standard, which includes a unique identifier for the digital record. As the SYNTHESYS+ project and the COST MOBILISE action have shown, setting up a mechanism for regular interactions within the community are facilitating these developments tremendously.

Objective 2: Engage non-technical users in the development process

Data standards often look complicated and very abstract to non-technical users. For most of the people involved in biodiversity research, it is not easy to detach the standard from the implementation of the standard. And in case an implementation of the standard doesn't exist, it is difficult to imagine how the data standard can potentially solve some of the problems with usage of that data.

During the SYNTHESYS+ project, we explored the use of a Wikibase instance to facilitate the inclusion of non-technical users. It provides an easy to use user interface, and it is also possible to have an early implementation of a standard. The project showed the potential of using Wikibase during the development, but it is definitely worth exploring this further in the future. It is a strategy that could be adopted by the TDWG working groups. One could even imagine integrating this approach more firmly into the TDWG development processes.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

However, SYNTHESYS+ also showed that it needs more to engage non-technical users. Easy to use forums where people can add use-cases and implementation needs. This can be through virtual meeting places (regular conference calls), virtual forums (e.g. GitHub issues) or in-person meetings. It is essential throughout the development process of a standard, that these modalities are supported and maintained.

Objective 3: Increase cross-pollination between disciplines through interoperability

Implementation of the IIIF standard for images of collection items, showed the possibility of making the images of natural science items available to other disciplines. By using this data standard, it becomes more straightforward to include them in other infrastructures such as Europeana. The reusability of the images is highly increased and it's allowing a broader range of stakeholders to make use of them.

In the future, it is definitely needed to keep an eye on the developments done in other (scientific) disciplines to ensure that we can maximally align the standards between each other. Massive amounts of data are becoming available through for example citizen science projects and satellite observations. In order to stay on top of this big data revolution, we need to make sure that our standards are aligned with these other disciplines.

Objective 4: To be able to report on the progress of digitization and improve the discoverability of specimens and collections

With the development of Latimer Core, the goal is to have a better aligned representation of the content of our collections. This includes the digitised parts of the collections, as well as the estimates of all collection items that are not yet digitally available. This is essential to be able to search for specific types of specimens, but also to guide policy makers towards the collections that are in need of digitisation efforts. Aligning the data on collections through the Latimer Core standard, allows the creation of overview dashboards such as the SYNTHESYS+ CDD. It therefore becomes crucial to the community that this standard is ratified as soon as possible, but also that registries of collections (such as the GBIF registry) are implementing this standard.

Not only the amount of specimens that are digitised is important to the community. Also the amount of data which is available on each of the specimens is crucial to steer funders towards the least digitised collections. Continuing the effort of developing a standard on minimal information on digital specimens (MIDS) will prove to be a powerful tool in mapping the out the digital completeness of collections.

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

Output generated by NA4

Over the course of SYNTHESYS+, networking action 4 generated other output that is not referenced specifically in this deliverable. For completeness, we are mentioning these publications here.

Dillen, M., Groom, Q., Agosti, D., Nielsen, L.H., 2019a. Zenodo, an Archive and Publishing Repository: A tale of two herbarium specimen pilot projects. BISS 3, e37080. https://doi.org/10.3897/biss.3.37080

Dillen, M., Groom, Q., Phillips, S., Spasic, I., 2019b. Next Steps in Data Capture from Specimen Labels and Data Integration: Lessons learnt from the ICEDIG pilots. BISS 3, e37081. https://doi.org/10.3897/biss.3.37081

Dillen M, Haston EM, Kearney N, Paul DL, Santos J, Shorthouse DP, Vaughan A, von Mering S, Groom Q (2021) Is Your Collection Ambiguous? Biodiversity Information Science and Standards 5: e73702. https://doi.org/10.3897/biss.5.73702

Fichtmueller D, Berendsohn W, Droege G, Glöckler F, Güntsch A, Hoffmann J, Holetschek J, Petersen M, Reimeier F (2019) ABCD 3.0 Ready to Use. Biodiversity Information Science and Standards 3: e37214. https://doi.org/10.3897/biss.3.37214

Groom Q, Bräuchler C, Cubey RWN, Dillen M, Huybrechts P, Kearney N, Klazenga N, Leachman S, Paul DL, Rogers H, Santos J, Shorthouse DP, Vaughan A, von Mering S, Haston EM (2022) The disambiguation of people names in biological collections. Biodiversity Data Journal 10: e86089. https://doi.org/10.3897/BDJ.10.e86089

Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., ... & Haston, E. (2020). People are essential to linking biodiversity data. Database, 2020. https://doi.org/10.1093/database/baaa072

Groom, Q. J., Dillen, M., Huybrechts, P., Johaadien, R., Kyriakopoulou, N., Fernandez, F. J. Q., … Wong, W. Y. (2021, March 3). Connecting molecular sequences to their voucher specimens. BioHackrXiv Preprints https://doi.org/10.37044/osf.io/93qf4

Hardisty AR, Addink W, Glöckler F, Güntsch A, Islam S, Weiland C (2021) A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). Research Ideas and Outcomes 7: e67379. https://doi.org/10.3897/rio.7.e67379

Petersen M, Hoffmann J, Glöckler F (2019) Access to Geosciences – Ways and Means to share and publish collection data. Research Ideas and Outcomes 5: e32987. https://doi.org/10.3897/rio.5.e32987

Schulman, L., Lahti, K., Piirainen, E. et al. The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. Sci Data 8, 137 (2021). https://doi.org/10.1038/s41597-021-00919-6

Semal, P., Adam, M., Van den Spiegel, D., Theeten, F., Engledow, H., Mergen, P., Gödderz, K., Rubio, A.C., 2019. CETAF Collection Dashboard: Mapping natural history collections diversity. BISS 3, e39667. https://doi.org/10.3897/biss.3.39667

Stoev, P., Haston, E.M., 2019. Progress in Authority Management of People Names for Collections. BISS 3, e35074. https://doi.org/10.3897/biss.3.35074

Mergen P, Trekels M, Leliaert F, Woodburn M, Droege G, Haston EM, Cubey RWN, Häffner E (2021) Describing Living Collections and Specimens . Biodiversity Information Science and Standards 5: e73697. https://doi.org/10.3897/biss.5.73697

Waagmeester, A., Willighagen, E. L., Su, A. I., Kutmon, M., Gayo, J. E. L., Fernández-Álvarez, D., ... & Koehorst, J. J. (2021). A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. BMC biology, 19(1), 1-14. https://doi.org/10.1186/s12915-020-00940-y

Woodburn M, Droege G, Grant S, Groom Q, Jones J, Trekels M, Vincent S, Webbink K (2021) A Data Standard for Dynamic Collection Descriptions. Biodiversity Information Science and Standards 5: e73902. https://doi.org/10.3897/biss.5.73902

References

References marked with an asterisk are direct output from SYNTHESYS+

Atlas, R. M. (2012). One Health: its origins and future. One health: The human-animal-environment interfaces in emerging infectious diseases, 1-13.

Baker, M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454 (2016). https://doi.org/10.1038/533452a

Blum S, Barker K, Baskauf S, Berendsohn W, Buttigieg P, Döring M, Droege G, Fichtmueller D, Glöckler F, Güntsch A, Guralnick R, Hoffmann J, Klazenga N, Macklin J, Morris P, Paul D, Petersen M, Robertson T, Sachs J, Shorthouse D, Walls R, Wieczorek J, Zermoglio P (2019) Integrating ABCD and DarwinCore: Toward a better foundation for biodiversity information standards. Biodiversity Information Science and Standards 3: e37491. https://doi.org/10.3897/biss.3.37491

Bisby, F. A. (2000). The quiet revolution: biodiversity informatics and the internet. Science, 289(5488), 2309-2312. https://doi.org/10.1126/science.289.5488.2309

Borgerud, C., Borglund, E. Open research data, an archival challenge?. Arch Sci 20, 279–302 (2020). https://doi.org/10.1007/s10502-020-09330-3

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

Clark, T., Martin, S., & Liefeld, T. (2004) Globally distributed object identification for biological knowledgebases, Briefings in Bioinformatics, Volume 5, Issue 1, Pages 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

Fichtmueller D. & Güntsch A. (2021). Milestone Report MS5.6 "A functional prototype of DiSSCo Modelling Framework". https://doi.org/10.34960/wn2h-9g16

Fichtmueller D. & Güntsch A. (2022). DiSSCo Prepare Deliverable Report D5.2 "DiSSCo Modelling Framework". https://doi.org/10.34960/e3nv-zh69

*Grant S, Jones J, Webbink K, Trekels M (2020) Reducing the Pain of Getting your Backlog Published. Biodiversity Information Science and Standards 4: e59183. https://doi.org/10.3897/biss.4.59183

Grieb J, Weiland C, Hardisty A, Addink W, Islam S, Younis S, Schmidt M (2021) Machine Learning as a Service for DiSSCo's Digital Specimen Architecture. Biodiversity Information Science and Standards 5: e75634. https://doi.org/10.3897/biss.5.75634

*Groom, Q., Dillen, M., Huybrechts, P., 2019a. Uniquely Identifying Collectors of Specimens. BISS 3, e37013. https://doi.org/10.3897/biss.3.37013

*Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., Page, R.D., Shorthouse, D.P., Thessen, A.E. and Haston, E. (2020) People are essential to linking biodiversity data, Database, 2020, baaa072, https://doi.org/10.1093/database/baaa072

*Groom, Q.J., Besombes, C., Brown, J., Chagnoux, S., Georgiev, T., Kearney, N., Marcer, A., Nicolson, N., Page, R., Phillips, S., Rainer, H., Riccardi, G., Röpert, D., Shorthouse, D.P. (2019b). Progress in Authority Management of People Names for Collections. Biodiversity Information Science and Standards, (7656). https://doi.org/10.3897/biss.3.35074

Groom, Q. & Poelen, J. (2020). Bat Interactions (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.3816676

Groom Q, Hyam R, Güntsch A, Stable identifiers for collection specimens. Nature 2017;546, 33. pmid:28569808 https://doi.org/10.1038/546033d

Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, et al., Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database. 2017(1): bax003. pmid:28365724 https://doi.org/10.1093/database/bax003

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

*Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, et al. (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLOS ONE 16(12): e0261130. https://doi.org/10.1371/journal.pone.0261130

*Hardisty A, Brack P, Goble C, Livermore L, Scott B, Groom Q, Owen S, Soiland-Reyes S; The Specimen Data Refinery: A Canonical Workflow Framework and FAIR Digital Object Approach to Speeding up Digital Mobilisation of Natural History Collections. Data Intelligence 2022; 4 (2): 320–341. doi: https://doi.org/10.1162/dint_a_00134

Hyam, R., Drinkwater, R. E., & Harris, D. J. (2012). Stable citations for herbarium specimens on the internet: an illustration from a taxonomic revision of Duboscia (Malvaceae). Phytotaxa, 73(1), 17-30. https://doi.org/10.11646/phytotaxa.73.1.4

Islam, S., Hardisty, A., Addink, W., Weiland, C. and Glöckler, F., 2020. Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. Data Science Journal, 19(1), p.50. DOI: http://doi.org/10.5334/dsj-2020-050

Klump, J., & Huber, R. (2017). 20 Years of Persistent Identifiers – Which Systems are Here to Stay?. Data Science Journal, 16, 9. DOI: http://doi.org/10.5334/dsj-2017-009

Leach, P.; Mealling, M.; Salz, R. (2005). A Universally Unique IDentifier (UUID) URN Namespace. Internet Engineering Task Force. RFC 4122. https://doi.org/10.17487/RFC4122

*Hyam, R (2019) Semantically linking specimens and images. Biodiversity Information Science and Standards 3: e35343. https://doi.org/10.3897/biss.3.35343

*Hyam, R (2021) Implementation of the IIIF for Natural History Collections. DiSSCo Knowledgebase. https://know.dissco.eu/handle/item/294

*Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, Röpert D, et al. (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLoS ONE 16(12): e0261130. https://doi.org/10.1371/journal.pone.0261130

Nelson, G. and Ellis, S. 2019 The history and impact of digitization and digital data mobilization on biodiversity researchPhil. Trans. R. Soc. B37420170391 20170391 http://doi.org/10.1098/rstb.2017.0391

Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K. OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. Publications. 2019; 7(2):38. https://doi.org/10.3390/publications7020038

*Smith VS, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston EM, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5: e46404. https://doi.org/10.3897/rio.5.e46404

SYNTHESYS+ was funded by the Horizon 2020 Framework of the European Union under the H2020 Open Innovation and Open Science Research Infrastructure call.

*Trekels M, Woodburn M, Paul DL, Grant S, Webbink K, Jones J, Groom Q (2020) How do you Develop a Data Standard? Wikibase might be the Solution…. Biodiversity Information Science and Standards 4: e59211. https://doi.org/10.3897/biss.4.59211

*Upham, N. S., Poelen, J. H., Paul, D., Groom, Q. J., Simmons, N. B., Vanhove, M. P., ... & Agosti, D. (2021). Liberating host–virus knowledge from biological dark data. The Lancet Planetary Health, 5(10), e746-e750. https://doi.org/10.1016/S2542-5196(21)00196-0

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Wittenburg, P., Anders, I., Blanchi, C., Buurman, M., Goble, C., Grieb, J., Hardisty, A., Islam, S., Jejkal, T., Kálmán, T., Kirkpatrick, C., Lannom, L., Lauer, T., Manepalli, G., Peters-von Gehlen, K., Pfeil, A., Quick, R., van de Sanden, M., Schwardmann, U., Soiland-Reyes, S., Stotzka, R., Trautt, Z., Van Uytvanck, D., Weiland, C., Wieder, P. (2022). FAIR Digital Object Demonstrators (2021). Zenodo. https://doi.org/10.5281/zenodo.5872645

*Woodburn M, Buschbom J, Droege G, Grant S, Groom Q, Jones J, Trekels M, Vincent S, Webbink K (2022) Latimer Core: A new data standard for collection descriptions. Biodiversity Information Science and Standards 6: e91159. https://doi.org/10.3897/biss.6.91159