H2020-INFRADEV-2019-2
Grant Agreement No 871043

Title
A best practice guide for semantic enhancement and improvement of semantic interoperability

Authors	Identifier of the author(s)
Mathias Dillen	https://orcid.org/0000-0002-3973-1252
Quentin Groom	https://orcid.org/0000-0002-0596-5376
Rob Cubey	https://orcid.org/0000-0001-7902-3843
Sabine von Mering	https://orcid.org/0000-0003-2982-7792
Alex Hardisty	https://orcid.org/0000-0002-0767-4310

Affiliation
Meise Botanic Garden, BE
Meise Botanic Garden, BE
Royal Botanic Garden Edinburgh, UK
Natural History Museum, Berlin, DE
Cardiff University, UK

Contributors
Josh Humphries
Ginger Butcher
Tim Robertson
Marcus Ernst
Sharif Islam

Publication year 2021

Creation date 06/01/2021

Funding Programme
H2020-INFRADEV-2019-2

Deliverable/milestone number
5.4

Dissemination level
Public

WP number WP5

Content keywords
technical

Project reference
DiSSCo Prepare (GA-871043)

Project output
Deliverable

Resource type
Text

Citation
Dillen et al. (2019) A best practice guide for semantic enhancement and improvement of semantic interoperability. DiSSCo Prepare deliverable 5.4 report. doi:10.34960/ajxs-zr25

Abstract

Textual data on natural history specimens regularly suffers from ambiguity and interoperability problems, which impairs their common understanding with related specimens and the connection of these properties to other sources of information. Semantic enhancement of these data is one approach to address these problems, where subjects and concepts are identified through common standards or links to authority resources rather than textual strings. Through enrichment, key properties of specimens such as (1) the location they were gathered from, (2) the agents who acted upon them and (3) their taxonomic determinations can be unambiguously identified and processed for further scientific research.

In this report, we take a closer look at the current state of natural history specimen data enrichment. Building on a range of pilot projects, we break down the general workflow of enrichment, informing on potential approaches/tools that may be utilized, key considerations that need to be made and obstacles that may be encountered. Workflows to enrich specimen data tend to be diverse, in particular because the context in which the enrichment takes place can be very variable. Not all institutions will have the same resources to undertake the enrichment process, nor are all collections managed and digitized in the same manner or can different types of collections be compared to each other.

Despite this lack of homogeneity, general lessons can be inferred and some base recommendations be stipulated. Enrichment should ideally take place in close accord with digitization. Otherwise, in general, enrichment will be easier to implement at larger scales above the collection level. Yet the key role in comprehensive enrichment of local knowledge about the collection and the relative ease at which low-hanging fruit can be (semi-)manually processed still promises considerable added value of even simple local approaches to enrichment. Technical obstacles, again, may be more easily tackled at big data levels, but this may lead to synchronization problems with local systems.

Data standards have adapted to support enriched properties in various manners. An extension for Darwin Core to accommodate agent attribution is under development and has been tested in this report. Some problems still abound, in particular the strain enrichment places on the simple data model of the popular Darwin Core archive. Alternative representations are gaining traction, including the openDS standard currently under development by DiSSCo.

Keywords

persistent identifier (PID), disambiguation, knowledge graph, darwin core, wikidata

H2020-INFRADEV-2019-2
Grant Agreement No 871043

Index

1. Introduction and definitions

In this report we document the process of enriching natural history specimen data with linked entities to define the semantics of their properties. Once enriched, specimens can be more easily discovered and connected to other, similar specimens. This facilitates the use of specimens for scientific research (Besnard et al. 2018), but also their curation by reducing the redundancy of digitization efforts and retrieving additional information through linked resources. Many properties of a specimen may be enriched, but in this report the focus lies on three general key properties of specimen data:

1) Agents, i.e. persons, groups or institutions associated with the specimen, such as who collected it in the field and who identified the taxon it belongs to.
2) Geographic features associated with the specimen, in particular the location where it was gathered.
3) The taxonomic name(s) given to the specimen.

Enrichment workflows may be implemented in different contexts and with different kinds of resources available. For example, enrichment may be done manually, through crowdsourcing initiatives or automatically using algorithms designed or trained for this purpose. Enrichment may also be performed at the local collection level or on an international scale, for instance as part of a service provided by the DiSSCo infrastructure. Different fields of natural history will have different traditions as to how specimens are documented after being gathered, not to mention variations in taxonomic conventions and preservation methods. This evidently influences any best practices that may be recommended. As a result, our recommendations are generalized to cover the enormous variety of specimens that are held in collections and will cover several scales and methodologies.

To support the recommendations made in this report, several case studies of past or ongoing work were investigated. Also, enrichment pilots were conducted by the contributors to this report and are documented here. Enrichment is not unique to the sector of natural history collections: key to it is the disambiguation of information, i.e. differentiating homonyms and equating synonyms. Disambiguation is a common problem in other fields of research, but the proportion of homonyms to synonyms may differ and, most importantly, the resources that should be consulted will differ. Natural history collections will also have their own peculiarities as a consequence of their close ties to the sciences of biology and geology, as well as to museums, archives and universities. Such peculiarities may facilitate or complicate enrichment, depending on their nature and the context.

In this first section, the main concepts will be defined and explored in more detail, drawing from work done in the ICEDIG (*Innovation and consolidation for large scale digitisation of natural heritage*) project - which preceded DiSSCo Prepare. In the second section, case studies will be described and enrichment pilots will be outlined, including both their methodologies and (initial) results. The third section breaks down the entire enrichment workflow, providing recommendations for best practices and the resources/technologies that may be employed. Finally, section 4 lies out the principal conclusions and recommendations for the future of enrichment of natural history specimens.

1.1. Semantic enhancement

The ICEDIG project was devised as the design study for the DiSSCo infrastructure. One of the key reports of ICEDIG was a blueprint for DiSSCo (Hardisty et al. 2020a), which was conceived as a synthesis of the most fundamental outcomes of the design study and includes many definitions of concepts and terms relevant for the constructing of DiSSCo. Among others, it defines a semantic assertion as "*The attachment (perhaps by reference to a defined vocabulary) of a specific meaning to a resource, attribute, property, etc*." Enhancement implies that we already have properties or attributes for specimens, but we somehow want to make them more meaningful. This is also often called (semantic) enrichment, as data become more usable (or 'richer') the more meaningful they are.

Meaning can be enhanced by making a data value more informative in relation to other, similar values. This can be done by applying a standard method of formatting information and by referring to authorities or other stable interpretative resources that are commonly understood. For example, a date can be and will be formatted in different ways, even if they all follow the Gregorian Calendar. As long as different formats are in use, it is difficult to understand what these dates mean without looking at them individually and having a human interpret them. Even then, ambiguity may exist, for instance when interpreting the day/month of "01/02/2020" or the century of "05-11-'89". As soon as a standard is commonly implemented, such as the ISO date standard (most recent version: ISO 8601-1:2019), dates can be easily aggregated and temporal analyses performed. Other quantitative values, such as geolocation and mass, benefit from common standards for units, reference systems and/or measurement protocols.

Other types of information may be less quantitative, such as the identities of people, taxonomic classifications of organisms and other scientific classifications for mineral samples, habitats or historic artifacts. These can be disambiguated by implementing unique identifiers and/or using common classification standards, like taxonomic backbones. Person names are commonly used to identify people, but names are often not unique and may be formatted and abbreviated in different ways. This can be described as the homonym and synonym problem, respectively (Deyun and Kayuzuki 2018). Organisms are regularly identified by their species, represented by a scientific name chosen according to the principles of binomial nomenclature. Scientific names date back centuries already and are a way to avoid various ambiguity problems with vernacular names, such as linguistic differences and cultural aspects. Scientific names mostly consist of Latin or Latinized words. They also reference their origin by incorporating the name of the person(s) first describing the species (as such).

Scientific names are intended to describe a group of organisms that are sufficiently similar to be considered to belong to the same taxonomic concept (taxon). Taxonomists propose names for specific taxa and these names are valid from the date of publication. The regulations for naming organisms are codified in the rules for biological nomenclature, notably the International Code of Nomenclature for algae, fungi, and plants (Turland et al. 2018) and the International Code of Zoological Nomenclature (ICZN 1999). These names are catalogued in nomenclatural registers such as IPNI, Zoobank, Mycobank and Phycobank. Scientific names can vary slightly in their spelling, but are largely stable once published.

However, this stability is certainly not true for the taxonomic concepts that these names refer to. Any organism can be identified to be a part of a certain taxonomic concept (i.e. species) and not of any other. Species themselves can be grouped together under higher taxon ranks such as genus, family and order. However, the criteria of delineating different taxa (such as species) and how they relate to other groups change over time depending on the opinion of taxonomic authorities and the availability of evidence. In recent years, the advent of molecular phylogenies has been an important driver of changes in taxonomy and shifting taxon concepts (Adamowicz 2015). This means that the meaning of scientific names can change over time when new insights, new research methods or new observations are taken into account. Therefore, while it is comparatively simple to semantically enhance specimen data with information on the scientific name given to it, persistently linking to the proper taxonomic concept is more problematic.

1.2. Semantic interoperability

Interoperability is one of the four principles of FAIR data: Findable, Accessible, Interoperable and Reusable (Wilkinson et al. 2016). In the ICEDIG report on interoperability of natural science collection data (Dillen et al. 2019), it is stated that the problem of semantic interoperability is “*the difficulty in integrating resources that were developed using different vocabularies and different perspectives on the data*” (as quoted from Heflin and Hendler 2000). This makes it difficult to have a common understanding of data properties and structure regardless of origin. Semantic enhancement is one way of improving interoperability, although it may still be hindered by a lack of common understanding of the attached meaning. Implementing data standards may not be effective if different, incompatible standards are in use or if considerable ambiguity of meaning still exists within the standard. Similarly, enrichment through external identifiers is only effective insofar as these identifiers are not ambiguous.

To address this lack of common understanding of attached meaning, a draft report from the FAIRsFAIR project on FAIR semantics (Hugo et al. 2020) provides guidance for creation and maintenance of what they call *semantic artefacts*. These are defined as "a machine-actionable and -readable formalisation of a conceptualisation enabling sharing and reuse by humans and machines." Controlled vocabularies, authority resources minting identifiers and data standards are all examples of such artefacts. As these artefacts can be in different formats and at varying levels of complexity, the FAIRsFAIR best practice guide can help the DiSSCo architecture and service design tasks to enable FAIR data and compliant services that fit the needs of the community. This includes services fundamental to semantic enhancement, providing persistent identifiers to unambiguously identify concepts such as persons, taxa or geographical features. However, as the FAIRsFAIR report is still in a draft stage and goes well beyond our scope of semantic enhancement, we will not go into further detail here. FAIR and the concept of semantic artefacts will be further explored in task 6.4.4.

Identifiers are commonly strings of text that represent a subject, concept or thing. In their most basic form, they are simple representations of the subject using a string, which is otherwise devoid of most or any meaning/connotation to the subject. The string may contain additional information or associations related to the subject, but doesn’t have to. A string may consist of numbers, alphanumeric codes or more complex constructs such as GUIDs (Globally Unique Identifiers). Due to their lack of implicit relation to their subject, identifier strings can more readily be kept unique and unambiguous even if information concerning the subject changes. Although essential for processing by machines, identifiers are not necessarily meant for vernacular use. If they are, they are commonly referred to as ‘names’.

Identifiers can be transient and break after whatever system keeping the connection between the identifier and the subject identified ceases to function properly. The connection may be lost or may be subject to change, for example due to taxonomic revisions as described in section 1.1. Identifiers may also not be globally unique, causing potential problems of ambiguity. To address these problems, the concept of Persistent Identifiers (‘PIDs’) has become popular in recent times (Hilse and Kothe 2006). Such identifiers are intended to offer a stable link with their subject in the long term and to be globally unique. However, the persistence of PIDs cannot be assumed to be a given. Rather, they are a promise of the organization minting the PID that the identifiers and their target subjects will be actively maintained to stay unique and reference the same subject.

Organizations minting PIDs are very useful for semantic enrichment approaches, as they absolve the enricher from minting and maintaining their own identifiers. Furthermore, if other resources make use of these identifiers, it becomes possible to further enrich specimen data by using the initial PID resource as a broker for other links. For example, enriching the name of a person who collected a certain botanical specimen with a Wikidata PID for this person enables the retrieval of other information linked through this Wikidata record, such as institutional affiliations, literary works they authored and geographic regions they have visited.

Wikidata is in fact a good example of how the use of semantic enhancement may facilitate the FAIRification of data:

- The use of PIDs renders the data more **findable**, as ambiguity is addressed.
- Wikidata supports multiple APIs (Application Programming Interface) that facilitate **accessibility**, both by machines and humans.
- By acting as a broker for other types of information or by resolving synonymic identifiers, the use of Wikidata identifiers improves **interoperability** with other data sources.
- Fundamental to Wikidata is a system of versioning. Keeping track of any change renders usage of these data replicable (i.e. **reusable**).

2. Case studies

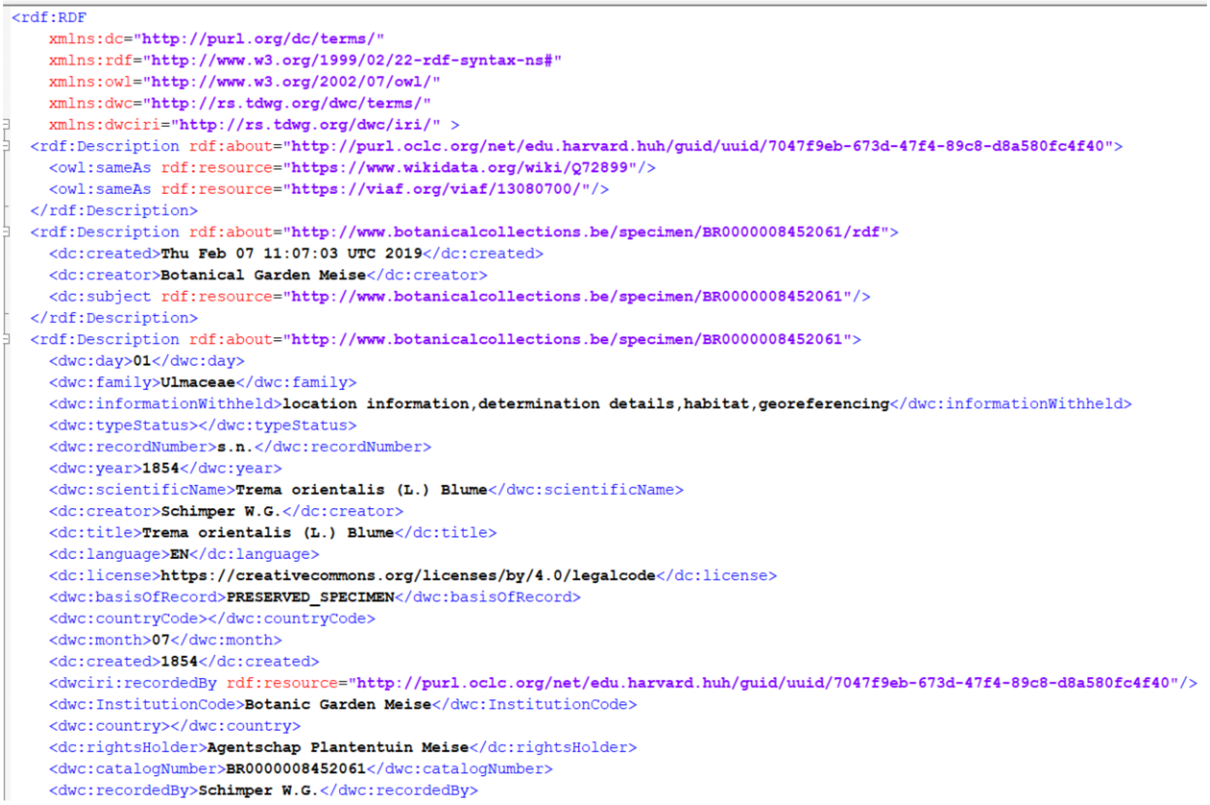
Over the last few years, there have been several semantic enrichment activities in the context of Natural History collections. In this section, we will briefly describe some of them and build on their achievements to inform on the best practices for an enrichment workflow in our context. Some of these activities consist of effective approaches to enrich specimen data, but there have also been developments in the realm of data standards on how enrichment is most optimally represented.

2.1. Botany Pilot

The Botany Pilot is an initiative led by the Botanic Garden and Botanical Museum Berlin (BGBM) and conceived in the Information Science & Technology Commission (ISTC) of the Consortium of European Taxonomic Facilities (CETAF). The aim of the Pilot is to link enriched specimen data from different herbaria in a SPARQL-queryable triple store. This way, specimens from different collections with properties in common (such as collector, taxon or location) can be connected to each other. Once this is done, identifications, transcriptions and georeferencing efforts done by one institution may propagate to others, reducing double-work and filling in gaps through complementary data.

The enriched content is provided by each contributor in the format of RDF/XML machine-readable renditions of their specimen data following Darwin Core terminology (Wieczorek et al. 2012), easily retrievable through the CETAF stable specimen identifier (Güntsch et al. 2017). Some example RDF/XML documents can be found below and in Fig. 1. Fig. 1 also illustrates how enrichment of persons is modeled in this document, making use of both Darwin Core and more general semantic web standards such as the Web Ontology Language (OWL). To improve interoperability between different contributors with CETAF identifiers, a minimal data standard has been conceived as to how specimen data are to be mapped onto Darwin Core in an RDF/XML format: the CETAF Specimen Preview Profile.

RDF examples:
https://www.botanicalcollections.be/specimen/BR0000014685156/rdf
https://herbarium.bgbm.org/data/rdf/B100165170



```
<rdf:RDF
    xmlns:dc="http://purl.org/dc/terms/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:dwciri="http://rs.tdwg.org/dwc/iri/" >
  <rdf:Description rdf:about="http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/7047f9eb-673d-47f4-89c8-d8a580fc4f40">
    <owl:sameAs rdf:resource="https://www.wikidata.org/wiki/Q72899"/>
    <owl:sameAs rdf:resource="https://viaf.org/viaf/13080700/"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.botanicalcollections.be/specimen/BR0000008452061/rdf">
    <dc:created>Thu Feb 07 11:07:03 UTC 2019</dc:created>
    <dc:creator>Botanical Garden Meise</dc:creator>
    <dc:subject rdf:resource="http://www.botanicalcollections.be/specimen/BR0000008452061"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.botanicalcollections.be/specimen/BR0000008452061">
    <dwc:day>01</dwc:day>
    <dwc:family>Ulmaceae</dwc:family>
    <dwc:informationWithheld>location information,determination details,habitat,georeferencing</dwc:informationWithheld>
    <dwc:typeStatus></dwc:typeStatus>
    <dwc:recordNumber>s.n.</dwc:recordNumber>
    <dwc:year>1854</dwc:year>
    <dwc:scientificName>Trema orientalis (L.) Blume</dwc:scientificName>
    <dc:creator>Schimper W.G.</dc:creator>
    <dc:title>Trema orientalis (L.) Blume</dc:title>
    <dc:language>EN</dc:language>
    <dc:license>https://creativecommons.org/licenses/by/4.0/legalcode</dc:license>
    <dwc:basisOfRecord>PRESERVED_SPECIMEN</dwc:basisOfRecord>
    <dwc:countryCode></dwc:countryCode>
    <dwc:month>07</dwc:month>
    <dc:created>1854</dc:created>
    <dwciri:recordedBy rdf:resource="http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/7047f9eb-673d-47f4-89c8-d8a580fc4f40"/>
    <dwc:InstitutionCode>Botanic Garden Meise</dwc:InstitutionCode>
    <dwc:country></dwc:country>
    <dc:rightsHolder>Agentschap Plantentuin Meise</dc:rightsHolder>
    <dwc:catalogNumber>BR0000008452061</dwc:catalogNumber>
    <dwc:recordedBy>Schimper W.G.</dwc:recordedBy>
```

Fig. 1: Example of specimen data in the RDF/XML format, following the CETAF Specimen Preview Profile specification. Note that the value for `dwc:recordedBy` *has been enriched as a URI under* `dwciri:recordedBy`*. Additional identifiers are added through an identity link (*`owl:sameAs`*) with the* `dwciri` *URI.*

Active contributors to the Pilot right now include the BGBM, Meise Botanic Garden (MeiseBG), The Royal Botanic Garden Edinburgh (RBGE) and the Natural History Museum of Vienna, as well as other institutions making use of the JACQ collection management system. The main condition for contribution is complying with the CETAF persistent identifier standard and achieving level 3. A dashboard listing the current levels for different CETAF institutions can be found here. The enriched content currently consists mainly of unique identifiers from multiple authority files (e.g. ORCID, VIAF) or brokers (e.g. Wikidata) for persons associated with specimens, such as collectors. BGBM also provides enriched geographic information in the form of Geonames IDs. More info can be found on the CETAF identifiers wiki. A paper describing the Pilot and its results in greater detail is in preparation.

An example of the triple store in action can be found here: Specimens from all contributing institutions are connected through a common property (their collector, Richard Spruce). In addition, by resolving the URIs for this person or using them in SPARQL queries, additional data on him can be easily retrieved from sources such as the Biodiversity Heritage Library (for literature) or Bionomia (for other specimens, see section 2.5).

2.2. COST Mobilise activities

2.2.1. Person identifiers workshops

The Mobilise COST Action aims to support progress in biodiversity informatics in Europe. In March 2019, Mobilise hosted a workshop on the Authority Management of People Names in Sofia, Bulgaria (https://osf.io/qwegk/wiki/home/). This workshop focused on the choices of identifiers used for people and was particularly important to establish Wikidata and ORCiD as important identifiers in our community. The results of this workshop were published in Groom et al. (2020).

Later, the Authority Management of People Names workshop was held as a pre-conference workshop of the Biodiversity Next Conference in Leiden in October 2019 (https://osf.io/9t3f2/). By using this venue, the organizers were able to attract a truly international audience to the discussion on the disambiguation of people's names and their links across digital infrastructure of biodiversity knowledge. The 24 attendees were divided into teams to work on different aspects of the problem, including analysis and visualization, data standards, disambiguation processes and engagement with the collections community. More details of the workshop can be found on the Open Science Framework site https://osf.io/9t3f2/wiki/home/.

In February 2020, a workshop focusing specifically on the topic of Wikidata was held in Warsaw. As an open linked database with a wide scope and a strong community, Wikidata had sparked increasing interest among the natural history collections community. This workshop was to serve as a starting point for different working groups addressing the different use cases of Wikidata for this community, the potential obstacles and the considerations that would have to be made. Particular attention was paid to Wikibase, the software technology used for Wikidata, which could also be implemented as novel instances of Wikidata with different scope and different community governance than Wikidata itself.

Four breakout groups were decided upon to cover different topics. One looked at the Big Picture, that is which use cases were realistic to implement within Wikidata and which were not, as well as the potential obstacles that would need to be overcome such as bias and overall management of these implementations. Another group dug more into the technical aspects, such as the overall limitations of the Wikidata infrastructure, the potential need for other Wikibases and how they could be integrated into our existing workflows. Finally, two groups looked at the data itself: one focusing on how taxonomy currently works in Wikidata, the problems with this model and how it could be amended; the other focusing on people, and by extension the organizations employing them, and their connection both to the data (such as specimens) they have collected and the literature they published. More info can be found on the workshop’s wiki.

2.2.2. Georeferencing workshop

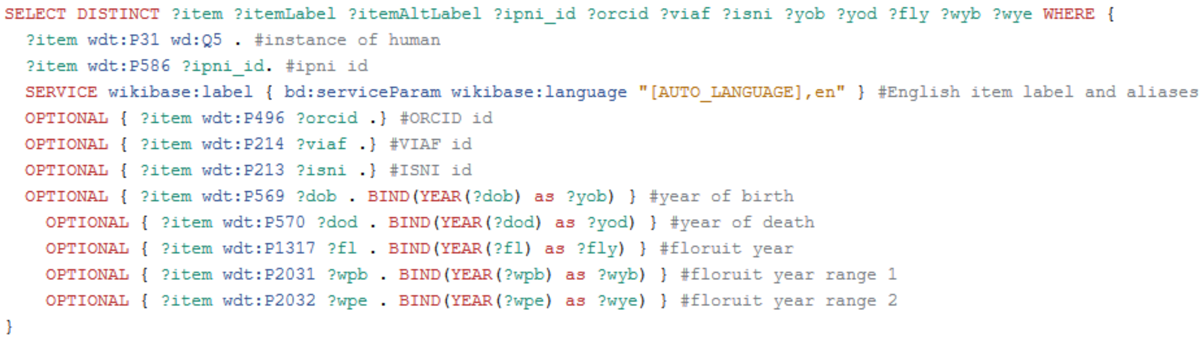
A workshop on the topic of georeferencing was held in February 2020 in Warsaw. The focus lied on investigating why georeferencing of natural history specimens is often still of poor quality. Various institutions presented their georeferencing efforts and the difficulties encountered along the way. Proposed causes for poor georeferencing included social, resource and technical reasons:

- General unawareness of the need for and importance of (proper) georeferencing.
- Technical bottlenecks in CMS and other databases, in particular quality control and compatibility with external services.
- Lots of double-work in a process that is relatively expensive and time consuming.
- Existing tools not optimized for effective use (by end-users or developers)
- Lack of good resources for enrichment

An extensive report of the workshop was published to Zenodo and can be found here. A peer-reviewed summary of the findings was also published (Marcer et al. 2020).

2.2.3. Automated person matching pilot

A Short Term Scientific Mission (STSM) in the COST Mobilise action was undertaken in early October 2019 by Mathias Dillen and hosted by Rod Page at the University of Glasgow. As part of this STSM, an R script was developed to automatically link strings of person names to Wikidata items, with or without temporal references such as date of birth or floruit date ranges (i.e. dates when these people were alive and working). The script obtained potential person records from Wikidata by combining six SPARQL queries (Fig. 2), which were all based on the presence of Wikidata properties that indicate the person being present in a natural history related authority source. The IDs used for the queries can be found in Table 1. In addition to these properties, generic person identifiers for ORCID, VIAF and ISNI were requested along with (English) Wikidata item labels as well as date of birth, date of death and floruit date properties. The queries can easily be replicated and are found in an Rmarkdown script on Github. All items were also subjected to the constraint of being instances of human.



```
SELECT DISTINCT ?item ?itemLabel ?itemAltLabel ?ipni_id ?orcid ?viaf ?isni ?yob ?yod ?fly ?wyb ?wye WHERE {
  ?item wdt:P31 wd:Q5 . #instance of human
  ?item wdt:P586 ?ipni_id. #ipni id
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" } #English item label and aliases
  OPTIONAL { ?item wdt:P496 ?orcid .} #ORCID id
  OPTIONAL { ?item wdt:P214 ?viaf .} #VIAF id
  OPTIONAL { ?item wdt:P213 ?isni .} #ISNI id
  OPTIONAL { ?item wdt:P569 ?dob . BIND(YEAR(?dob) as ?yob) } #year of birth
    OPTIONAL { ?item wdt:P570 ?dod . BIND(YEAR(?dod) as ?yod) } #year of death
    OPTIONAL { ?item wdt:P1317 ?fl . BIND(YEAR(?fl) as ?fly) } #floruit year
    OPTIONAL { ?item wdt:P2031 ?wpb . BIND(YEAR(?wpb) as ?wyb) } #floruit year range 1
    OPTIONAL { ?item wdt:P2032 ?wpe . BIND(YEAR(?wpe) as ?wye) } #floruit year range 2
}
```

Fig. 2: One of the SPARQL queries (for IPNI author id).

Wikidata Property	ID	% of unique items
IPNI author ID	P586	51
Harvard Index of Botanists ID	P6264	21
Entomologists of the World ID	P5370	10
Zoobank author ID	P2006	15
BHL creator ID	P4081	23
Wikispecies (not a property)		43
VIAF ID	P214	30
ISNI ID	P213	23
ORCID	P496	3

Table 1: Wikidata properties used for the SPARQL queries and the different resources they are referencing. Also indicated is the percentage of all Wikidata items found this way which had this property. Wikispecies has no specific property, but can be queried using its Wikimedia schema. VIAF, ISNI and ORCID were considered too generic and were only part of the results for the other properties.

Jointly, these six queries obtained 104.625 Wikidata items during the latest run on 2020-04-08. Almost half of these items did not have a date of birth and few had floruit dates. The presence of the identifier property is indicated in table 1. There was also considerable overlap: in particular for Zoobank and the Harvard Index of Botanists, of which almost all persons were also present in Wikispecies and IPNI respectively.

To increase the amount of date information for these Wikidata items, an attempt was made to deduce floruit dates based on publications in Wikidata authored by these persons. For this, a series of SPARQL queries were performed for each person, finding all publications listing them as an author. However, this attempted enrichment of Wikidata properties had only a marginal impact, as many publications in Wikidata lacked enriched author information and/or lacked a publication date.

The actual matching of a source dataset of people associated with specimens (e.g. collectors) to the set of Wikidata items follows a rule-based workflow which employs date filters and fuzzy string matching. A flowchart can be found in Fig. 3 and the workflow is also described in text in the two boxes below. Box 1 describes the overall workflow visualized in Fig. 3. Box 2 specifies the fuzzy matching approach. Each person record is compared to the Wikidata set and all the matched items are withheld as results.

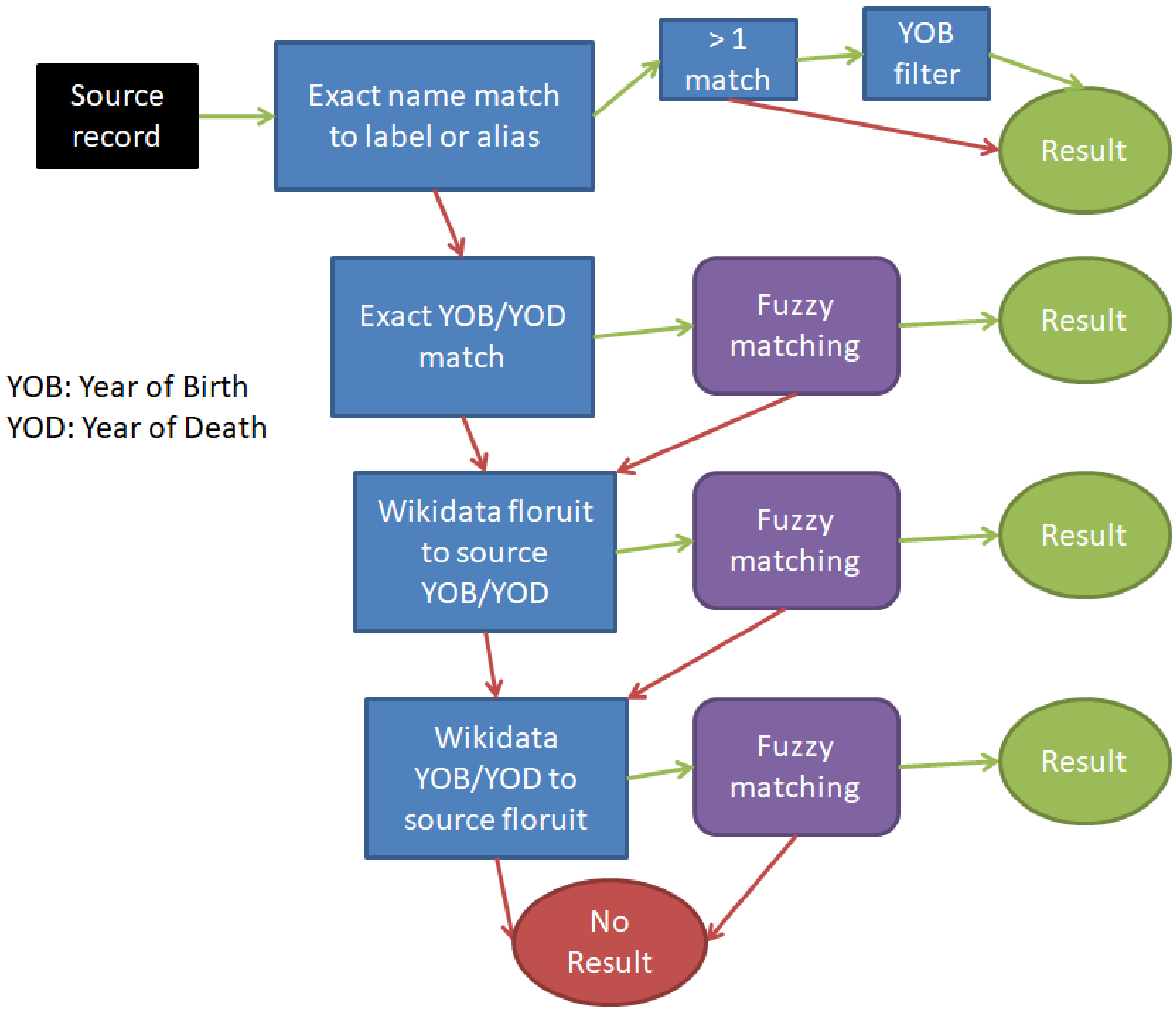


Fig. 3: Flowchart of the matching process (see also Box 1). Each unique collector record is first subjected to exact matching of name to Wikidata item label, followed by fuzzy matching (see Box 2) after filtering the Wikidata candidate items based on three sequential date conditions.

Box 1: Matching workflow

For each name in the source dataset:

1) Find all exact string matches of the full name to the Wikidata label/aliases.

    1a) If multiple results, additional filter for year of birth (if any).

    1b) If any result, print and next.

2) If no exact match was found:

    2a) Filter on year of birth or death exact match

2b) If any yob/yod matches found, fuzzy matching. If any results of this matching, print and next.

3) If no exact yob/yod matches found:

3a) Filter on floruit date 1 > yob + 15 and floruit date 2 < yod

3b) If any matches and birth/death dates in source were available, fuzzy matching. If any results of this matching, print and next.

4) Else and if there is a minimal floruit date in source:

4a) Filter that there is a yob or yod in Wikidata.

4b) Filter that either or both fit in the floruit range (-15y for childhood).

4c) If any matches, fuzzy matching.

There was no check for floruit matches of source into floruit matches of Wikidata.

Box 2: Fuzzy matching process:

1) Fuzzy match a last name into the (presumed) last name of the Wikidata items.

2) Remove records where the length of the source last name is more than 1 char longer.

3) If first name available:

3a) If the first name contains dots, try to exact match to the initials of the Wikidata name (both processed to remove the dots).

3b) Otherwise, fuzzy match the first name into the Wikidata full name.

4) If middle name available:

4a) Fuzzy match it into the full Wikidata name

5) Return all results still withheld.

As an example dataset, the known collector names of MeiseBG's herbarium were processed using this script. These collectors are curated in a separate table of MeiseBG's collection management system (BG-Base) and tally at around 6.500 different names. All of them have a (family) name. Many have first names and some have middle names in separate fields as well. Few have dates of birth or dates of death. By connection to the specimens they collected, floruit date ranges could be inferred from specimen collection dates. However, a few of these collection dates (or collector connections) are incorrect. Therefore, inferred floruit ranges larger than 100 years were excluded.

As a result of processing the names of MeiseBG collectors, 2.599 persons could be linked to a single Wikidata item. Almost 60% of these were due to exact matching between the collector's full name and Wikidata's item label. 3.300 could not be matched at all. The remaining 10% required further disambiguation. The fuzzy matching approach had an error rate of 4-6%, based on a manual validation process: Non-exact single matches were estimated by a validator to be either correct, wrong or suspicious. Matches were labeled as suspicious if it was not immediately clear whether the collector and the Wikidata item referred to the same person - these make out the uncertainty between the 4 and 6%. This validation process does not address ambiguous multi-matches or false positive exact matches.

To address this error rate, a follow-up validation step was put in place. In this step, all connected identifiers were cross-checked with a separate dataset of matched collector names and identifiers. This set was assembled through a semi-automated process making use of the built-in cluster and fuzzy matching algorithms of OpenRefine. This workflow was more time-consuming than the automated approach described above and more difficult to document or replicate. By considering the OpenRefine method a relatively independent approach, a consensus set of matched collector names was derived based on two criteria

(1) The match was withheld if both methods connected the name to at least one identical identifier (e.g. both methods connect a name to the same IPNI author ID)
(2) It was also withheld if there was no overlap in identifier source (e.g. one method matched a name to Zoobank, whereas the other method did not consider Zoobank as a source)

As an outcome of applying this script and validation of the results to MeiseBG collector information, the number of enriched collectors was vastly increased. Earlier, about 700 collectors had been enriched manually by an expert. With the additional enrichment done by the script, this number was brought up to 1.724, covering more than 60% of specimens within the collection that had a known collector. The enriched information was made available on the institutional portal, where it could be harvested from the XML/RDF for the Botany Pilot (section 2.1), as a concrete example of a use case. It was also made available on GBIF, insofar as the Darwin Core schema allowed it. For more information on the GBIF publication, see sections 2.3 and 2.4.

This matching process has also been adapted to Python in Jupyter notebooks by Niels Klazenga of Royal Botanic Gardens Victoria, who made use of n grams clusters rather than line-by-line fuzzy string matching. This has the advantage of being computationally much more efficient. He also opted to omit any date filters and use dates for post-hoc validation instead.

2.3. Agent Attribution task group

An Attribution interest group has been set up as a collaboration between the Biodiversity Information Standards organization (TDWG) and the Research Data Alliance (RDA). In early 2020, a task group started within this interest group on the topic of People in Biodiversity Data. This task group is working on a Darwin Core Agents Attribution extension, which allows the identification (by string or identifier) of people related to a specimen by some action they performed on it (gathering, identifying, transcribing, mounting, etc.). The extension is documented and can be further discussed on Github. Table 2 lists the current set of terms and their definitions. Table 3 lists the current recommended controlled vocabularies.

term	definition	vocabulary
`occurrenceID`	ID of the occurrence the agent acted upon.	
`agentType`	The nature of the agent.	yes
`agentIdentifierType`	The type of identifier for the agent.	yes
`identifier`	A string conforming to an identification system.	
`name`	The name of the agent.	
`alternateName`	An alias for the item. Other full name agent may have been known under such as maiden name.	
`verbatimName`	As written on occurrence, such as the collection or determination label.	
`action`	The name of the single action written as a verb in past tense.	yes
`role`	The name of the role the agent played in the context of executing the action.	yes
`displayOrder`	The display order for the agent that executed the action when more than one agent was a participant.	integer
`identificationID`	An identifier for the Identification, i.e. the body of information associated with the assignment of a scientific name.	
`startedAtTime`	Start is when an action is deemed to have been started by an agent.	ISO date
`endedAtTime`	End is when an action is deemed to have been ended by an agent.	ISO date

Table 2: List of terms used in the most recent draft of the Agent Attribution extension (as of 2020-12-14). Term definitions are (briefly) indicated, as well as whether a controlled vocabulary or any other restriction exists.

agentType	agentIdentifierType	action	role
Person	*ORCID*	*collected*	*specimen collection role*
Organization	*VIAF*	*identified*	*primary collector role*
Software-Application	*ISNI*	*verified*	
	ResearcherID	*observed*	
	HUH	*prepared*	
	GRID	*preserved*	
	ringgold	*georeferenced*	
	RoR	*measured*	
	wikidata	*transcribed*	

Table 3: List of recommended controlled vocabularies for some of the terms (as of 2020-12-14).

Implementation of this extension can now be tested using the GBIF Integrated Publishing Toolkit (IPT) in test mode. An example specimen record with this implementation from the MeiseBG herbarium dataset can be found here in a JSON format. A backup of this JSON file is available as Example 1 in appendix 6.1, as the GBIF test environment is periodically refreshed, and therefore not considered suitable for long term linking.

This specimen was collected by two persons. For one of them, two different persistent identifiers are known (an ORCID and a Wikidata ID). For both, a string with their individual name is listed separately (`dwc:name`) as well as the order in which they appear on the specimen's label (`dwc:displayOrder`). The literal way their names were rendered on the specimen label is mapped to `dwc:verbatimName`. For `dwc:alternateName`, the label under which this team is known in the source database was used.

Given the star schema structure of Darwin Core, the number of records in this extension can increase rapidly. For ca. 1.7M specimen records, the extension totaled to ca. 5.2M records. This happens because there may be multiple identifiers, for multiple individuals part of a team, all repeated for each specimen they are linked with. Specimens may also be associated with multiple identifications (species determinations) done by different agents. These numbers may increase even more if other actions than collecting and identifying are considered as well.

Here (and example 2 in Appendix 6.1) is another example with a team of 4 members. Only one has any PIDs, but all members are listed separately with their display order and their individual names in the `dwc:name` field. This should facilitate any future enrichment activities.

Finally, an example (example 3 in Appendix 6.1) of how inflated the extension can get: this record has 52 agent attribution records for a single specimen. There are only two persons associated with the specimen, but they have 8 and 9 associated identifiers respectively. One was the collector who also initially identified it. The other added identifications at four different dates.

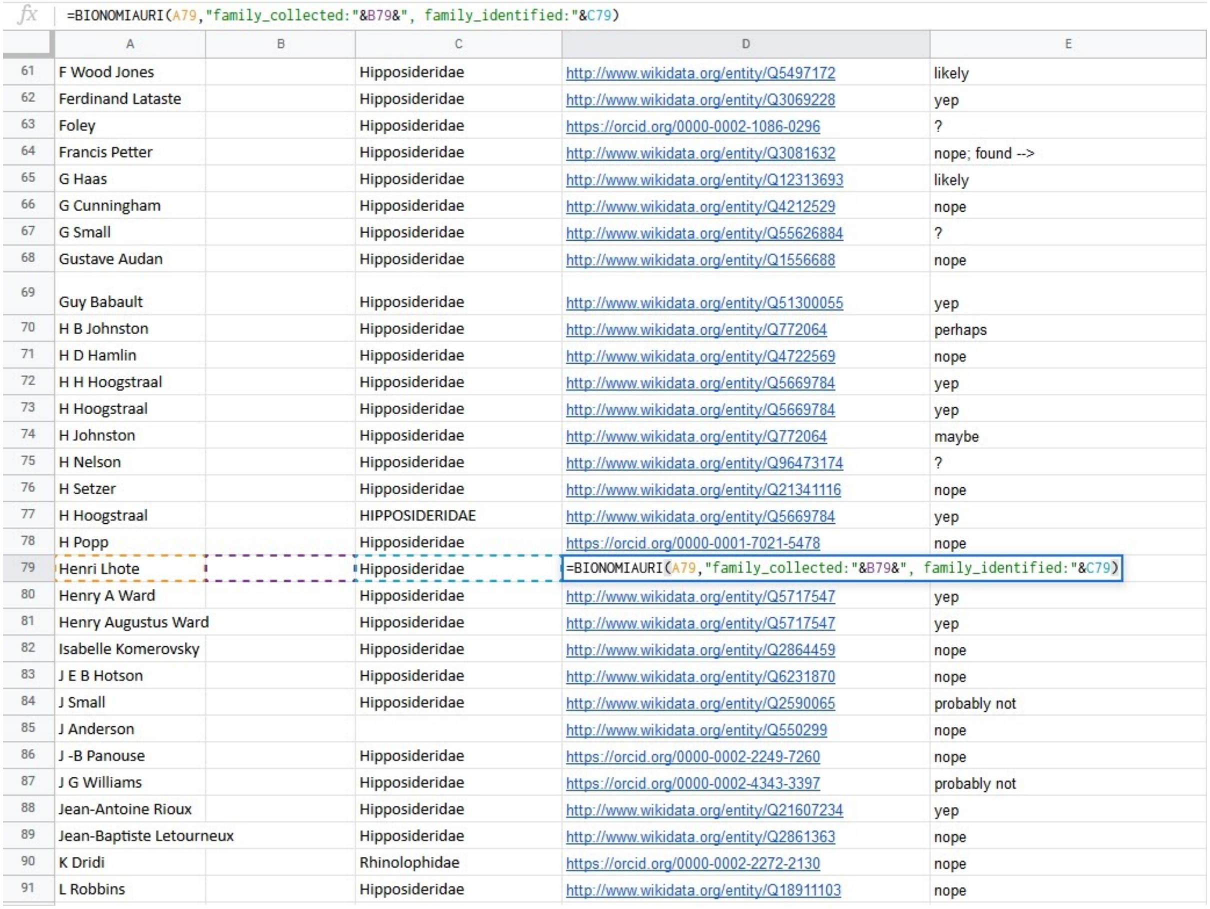
2.4. GBIF person ID terms

In June 2020, after discussion in the Attribution task group (see section 2.3), GBIF added two new terms called `gbif:recordedByID` and `gbif:identifiedByID` to its DwC Occurrence Core schema. These terms are intended for URIs identifying persons or other agents who would otherwise be put as text in the `dwc:recordedBy` and `dwc:identifiedBy` fields. As they only occur in the core table, only one URI can be added to an occurrence record for each of the two new terms. Due to this development, ORCIDs are taken now from observations recorded using iNaturalist and IDs can also be supplied by other providers, for instance through IPT installations. An example dataset making use of these new terms is Meise Botanic Garden (2020).

2.5. Bionomia

Bionomia is a web tool developed and maintained by David Shorthouse, which uses crowdsourcing to attribute specimens to their collectors. It harvests `dwc:recordedBy` and `dwc:identifiedBy` data from GBIF and facilitates users to connect these person name strings to ORCIDs (for living people) or Wikidata URIs (for the deceased). Hence, specimen data with these person name strings can be connected to ORCIDs or Wikidata items.

An API is available and a Google Sheets plugin has been developed relying on this API to automatically parse name strings and present likely ORCIDs/Wikidata URIs (Fig. 4). Additional disambiguating information can be added to the request, such as taxonomic family names the person is known to have worked with. A plugin is also available for Chrome and Firefox to automatically render these identifiers on GBIF specimen occurrence pages.



```
=BIONOMIAURI(A79,"family_collected:"&B79&", family_identified:"&C79)
```

	A	B	C	D	E
61	F Wood Jones		Hipposideridae	http://www.wikidata.org/entity/Q5497172	likely
62	Ferdinand Lataste		Hipposideridae	http://www.wikidata.org/entity/Q3069228	yep
63	Foley		Hipposideridae	https://orcid.org/0000-0002-1086-0296	?
64	Francis Petter		Hipposideridae	http://www.wikidata.org/entity/Q3081632	nope; found -->
65	G Haas		Hipposideridae	http://www.wikidata.org/entity/Q12313693	likely
66	G Cunningham		Hipposideridae	http://www.wikidata.org/entity/Q4212529	nope
67	G Small		Hipposideridae	http://www.wikidata.org/entity/Q55626884	?
68	Gustave Audan		Hipposideridae	http://www.wikidata.org/entity/Q1556688	nope
69	Guy Babault		Hipposideridae	http://www.wikidata.org/entity/Q51300055	yep
70	H B Johnston		Hipposideridae	http://www.wikidata.org/entity/Q772064	perhaps
71	H D Hamlin		Hipposideridae	http://www.wikidata.org/entity/Q4722569	nope
72	H H Hoogstraal		Hipposideridae	http://www.wikidata.org/entity/Q5669784	yep
73	H Hoogstraal		Hipposideridae	http://www.wikidata.org/entity/Q5669784	yep
74	H Johnston		Hipposideridae	http://www.wikidata.org/entity/Q772064	maybe
75	H Nelson		Hipposideridae	http://www.wikidata.org/entity/Q96473174	?
76	H Setzer		Hipposideridae	http://www.wikidata.org/entity/Q21341116	nope
77	H Hoogstraal		HIPPOSIDERIDAE	http://www.wikidata.org/entity/Q5669784	yep
78	H Popp		Hipposideridae	https://orcid.org/0000-0001-7021-5478	nope
79	Henri Lhote		Hipposideridae	=BIONOMIAURI(A79,"family_collected:"&B79&", family_identified:"&C79)	
80	Henry A Ward		Hipposideridae	http://www.wikidata.org/entity/Q5717547	yep
81	Henry Augustus Ward		Hipposideridae	http://www.wikidata.org/entity/Q5717547	yep
82	Isabelle Komerovsky		Hipposideridae	http://www.wikidata.org/entity/Q2864459	nope
83	J E B Hotson		Hipposideridae	http://www.wikidata.org/entity/Q6231870	nope
84	J Small		Hipposideridae	http://www.wikidata.org/entity/Q2590065	probably not
85	J Anderson			http://www.wikidata.org/entity/Q550299	nope
86	J -B Panouse		Hipposideridae	https://orcid.org/0000-0002-2249-7260	nope
87	J G Williams		Hipposideridae	https://orcid.org/0000-0002-4343-3397	probably not
88	Jean-Antoine Rioux		Hipposideridae	http://www.wikidata.org/entity/Q21607234	yep
89	Jean-Baptiste Letourneux		Hipposideridae	http://www.wikidata.org/entity/Q2861363	nope
90	K Dridi		Rhinolophidae	https://orcid.org/0000-0002-2272-2130	nope
91	L Robbins		Hipposideridae	http://www.wikidata.org/entity/Q18911103	nope

Fig. 4: Example of the Google Sheets plugin in action.

See also “Natural History Collections Data Roundtrip: GBIF, Wikidata, Bloodhound, ORCID and back again” https://www.youtube.com/watch?v=SCdDIDfDngc , Online Seminar, 2020-04-06, 1 h 8 min

2.6. Geonames pilots

Inspired by Geonames enrichment work at BGBM, a matching process was constructed for MeiseBG specimen data, working with a Geonames data dump (allCountries.zip, taken at 2020-09-09). For the matching process, all unique locality strings from MeiseBG’s published herbarium collection were listed (549.591 total). The following methodology was employed, of which the R code can be found here:

1) The herbarium specimen locality strings were split up by several punctuate delimiters: , ; - : (/ ' ". The “-” delimiter had to be preceded by a space, to avoid its usage as a hyphen.

2) The alphabetic characters were extracted from each substring split this way.

3) A matching process to Geonames records was set up on a country-by-country basis, using the ISO country code as a key. Some specimen records were omitted this way, either as no country code was known for them or their code was incompatible with the codes used by Geonames.

4) For each country, all labels of Geonames records - both names and alternate names - were extracted and their alphabetic characters exactly matched to the similarly extracted texts from the Meise specimen data. Doing the matching per country makes the process more efficient and avoids some homonym problems.

5) For validation, ambiguous matches were excluded: this includes multiple Geonames records for a single locality substring and different Geonames records for different substrings of one locality string. For example, the locality string “Bokunu (Unatra)” matches two Geonames records: one for Bokunu and one for Unatra. Alternatively the locality string “Kipako” matches three different Geonames records with the exact same label.

6) Because the matching was simplified to alphabetical characters only, some false positives can occur for place names with other characters (typically accents). As a quick fix for this, matches based on substrings with length less than 4 were excluded as well. This also takes care of spurious matches for indications of distance, height or wind directions. It will, however, also exclude some correct matches.

7) In an additional validation step, coordinate data of georeferenced specimens was used to validate the matching results. If multiple specimens shared the same locality, but had different coordinates, the minimal difference with the Geonames coordinates was assumed - as specimens may be georeferenced incorrectly. A maximum difference between the decimal coordinates of the specimen and of Geonames of 2 was permitted. This is quite loose, but localities regularly indicate provinces which can cover massive distances.

Through this process, 85.841 different locality strings could be connected to single Geonames records. Much more links could be made if the process were refined further or through more post-processing of the results, as 196.515 locality strings were not considered because they were matched to multiple Geonames records. An additional 9.789 localities had different Geonames matches for different substrings.

2.7. Dataset authorship attribution

The Natural History Museum London's Data Portal (https://data.nhm.ac.uk) hosts the Museum's digitised specimen collections as well as datasets produced by its scientists in support of their research publications. Much like other data repositories, the Data Portal stores metadata with each of these research datasets to capture important information such as licensing and authorship.

Currently, the authors of each dataset are captured as a list of strings which is problematic for many reasons and represents a very simple attribution model. To improve this system and create richer connections between Data Portal datasets and other systems through the DOIs the Data Portal mints for each dataset, a new data model is presented (and will be implemented in due course) which replaces these strings with a proper attribution model. This model provides a significant improvement over the existing system, allowing modelling of different attribution activities, agents and roles with the possibility for extensions in the future. Furthermore, once the model is mapped to the DataCite DOI metadata standard, the Data Portal will be able to produce richer DOIs with linkage to external systems like ORCID and ROR.

The model has been designed by taking guidance from the joint RDA/TDWG attribution recommendations (Thessen et al., 2019) and DataCite's DOI metadata schema (DataCite Metadata Working Group, 2018). The RDA/TDWG recommendations are particularly applicable given the types of data the Data Portal hosts and the DataCite DOI schema is referenced as minting DOIs is one of the primary outputs of the Data Portal and therefore any new attribution model must eventually be mappable to this schema. The Data Portal mints DOIs for each dataset using DataCite and includes author information in the DOI's metadata. It is important to note that the real world limitations and requirements of the Data Portal's existing models and requirements influence the design of this attribution model heavily.

In the below schema diagram (Fig. 5), the "Package" represents a dataset in Data Portal terminology and therefore the "Package" concept is analogous to the "Entity" in the RDA/TDWG recommended PROV terminology (Belhajjame et al., 2013). The author field in the Package table is part of the existing attribution method and is not used in this new attribution model but will continue to exist as it is part of the core CKAN (https://ckan.org/) model on which the Data Portal is built.

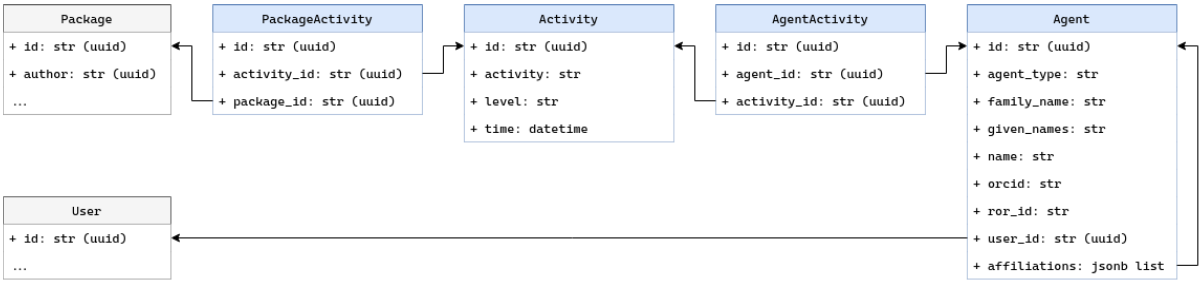


Fig. 5: Schema of the dataset attribution model. Blue tables represent new tables and grey tables are the existing Data Portal tables.

As per the PROV model, the Agent table represents either a person or an organisation (via the agent_type field). If the Agent is a person, they can have an associated ORCID, an associated user_id which references a standard Data Portal user and an associated set of affiliations to other Agents (for example, to represent a person belonging to a particular organisation). If the Agent is an organisation, they can have an associated ROR identifier.

The Activity and the two join tables link the Packages (the Entities) and the Agents with an Activity that the Agent has performed on the Package (Entity). The vocabulary used for the activity field isn't finalised yet but is likely to be a mix of and DataCite/CRediT (http://credit.niso.org/) terminology. The vocabulary will be catered to fitting the DataCite DOI schema and other Data Portal requirements. The level field is included as part of mapping the model to CRediT.

This model for attribution is initially focused on supporting typical digital curation activities on the Data Portal, like people "creating" datasets and others "contributing" to them. However, alignment with the RDA attribution model means that the data model is not restrictive and allows for future work to extend the entities, activities and agents beyond this scope. A later focus will be to roll this out to attributing physical and digital curation activities at a specimen level, including preparation, georeferencing, transcribing and other steps in the Museum's digitisation workflows. These data can then be presented in the Data Portal's specimen record interface, and through the DwC-A the Data Portal already generates, in alignment with the new Darwin Core extension for attribution. There are currently limitations to the information the Museum's collections management system captures that can be used to generate granular attribution data at a specimen level, and this is also a factor that will need to be addressed as part of the longer term plans.

3. The enrichment workflow

The aim is to enrich the digital representations of physical collection specimens so that the data on them are more meaningful, that is to say, it can be more easily understood by other users and by machines. The priority fields we focus on in this report are people, taxa and geographic features (who, what and where). These are key data to making specimens findable and usable for science (Groom et al. 2019). There are numerous factors to consider when building an enrichment workflow and we will discuss them separately below.

It is important to first have a working knowledge of what sort of data are to be enriched (3.1) and a clear picture of what the enriched product will look like (3.2). Both of these factors will also be influenced by the context in which the enrichment takes place (3.3), such as the scale, the level of achieved digitization and the resources available for implementing the workflow. These factors will also inform the approach chosen for the effective enrichment steps (3.4 and 3.5). Finally, there are considerations to be made as to how enriched data get published and how they can be kept synchronized, up-to-date or at the very least resolving (3.6 and 3.7).

3.1. What kind of data?

Digital records for physical specimens may contain information referring to people, location and/or taxonomy to various extent. Multiple persons may be associated with a physical specimen, either performing different actions on it or responsible for the same action, possibly at different times (such as different taxonomic identifications). Examples of different actions can be found in the draft Darwin Core Agent Attribution Extension.

If the identity of a person is unknown or unclear, other information known about the specimen can be used to retrieve or refine this identity. The point where the specimen was collected in space and time is useful in this regard, but other clues may be just as telling. These include the use of words, phrasing and syntax that can be associated with a person, but also institutional affiliations, the structure of the field number and the person's handwriting. Even the nature of the specimen itself, as most collectors have preferences for taxa and habitats.

Similarly to people, multiple taxon names may be associated with a specimen. A specimen can be a nomenclatural type for multiple taxon names (Turland et al., 2018). Different opinions on the identification of the specimen may have been recorded. Not all of these taxon names are equally important, as some may be considered to be more probable and others to be effectively incorrect. If a taxon name is missing or unclear, it can also be inferred from other pieces of information. However, the assumption that specimens collected in the same place and at the same time, or described in very similar ways, are of the same species or even higher taxon group will not always be correct. This approach can be refined or even replaced by applying Machine Learning algorithms to an image of the specimen in order to recognize a species (or taxon) based on the organism's physical characteristics (Carranza-Rojas et al. 2017, Little et al. 2020). These algorithms are still being developed and refined, so there is merit in combining both approaches.

For geographic features, we first need to establish what exactly is meant by this. Most often, the origin of a specimen is of interest, i.e. the location where it was collected. Other geographic features may also be associated with a specimen:

- A substrate, which can be another organism (a tree, a host).
- An area. Some species may be mobile and the wider area is more interesting than the specific coordinates. This can also be a geographic feature that implies additional constraints, such as a cave system, a river or a beach. Areas may also be more complex than can be construed from simple coordinate points, such as in the case of valleys, mountains or lakes.
- A habitat. Habitats may provide additional information about the specimen on top of its exact gathering location. This includes the habitat *at the time*, which may have changed since, but also the state of the ecosystem at the time of gathering for dynamic systems (like mudflats). Multiple classification systems exist for habitats, such as the European EUNIS or the IUCN Habitats Classification Scheme.
- Remarks by the collector. This type of information can provide hints or effective content on any of the above features. It can also provide other hints, such as descriptions of weather or details about the collecting trip.

Other geographical features not directly related to the specimen's collection can be noted:

- Specimens can be collected alive and raised in another location, before ultimately ending up (partially or completely) in a preserved collection. It can be of interest to enrich both the location where the specimen was initially collected and where it was grown before harvest/death.
- Locations where the specimen has been kept over time. These can be institutional affiliations.

3.2. What is our target?

For the three properties covered here, only taxon names have a data standard that may be suitable for our enrichment goal, i.e. the rules of binomial nomenclature. Even there, issues can come up, such as synonyms, spelling mistakes or punctuation differences. Hence, why it still can be more useful to work with identifiers such as those minted in the Catalogue of Life or the World Flora Online (Roskov et al. 2020, WFO 2019). Alternatively, taxon resolution services are already available to fix a lot of these issues automatically, like the Taxonomic Name Resolution Service (Boyle et al. 2013). Services also exist to split taxon name strings into their individual parts, e.g. gnparser. Problems concerning synonyms and conflicts (such as homonyms) are best addressed at a more central level than the collection level, where the scale of the problem is much more apparent and a wider expertise can be consulted. Jin and Yang (2020) describe a workflow to clean up common taxonomic and geographic errors for large datasets of biodiversity occurrences.

Geographic features can be identified using point coordinates, a reference system (CRS) and radial uncertainty (Chapman and Wieczorek 2020). However, uncertainty may not be radial, as for instance when georeferencing is inferred from survey grids such as the Belgian IFBL or the British Ordnance Survey National Grid. For more complex features, such as rivers or areas, which cannot be approximated by circles, other shapes than points can be used, such as lines, polygons or even three dimensional shapes when elevation or depth is important. It is also possible to delegate the exact spatial description to an authority source, such as Geonames or the Getty Thesaurus of Geographic Names. In this case, a feature would be enriched by a unique identifier (PID) for it, minted by the authority source. The precision of such identifiers may still vary, similar to how the precision of place names is variable: locality descriptions can consist of names for large provinces, small towns or even elaborate directions to specific landmarks.

For person names, standards exist, such as the Chicago Manual of Style. In certain fields, such as academic publishing, author names can be styled in different ways depending on the citation system in use, e.g. of the American Psychological Association (APA 2020) or the Modern Language Association (MLA 2017). Database systems may have their own parsing of names, such as separate fields for first name, family name, title, prefix and much more. Complications may arise between different cultures, incomplete names, changed names (e.g. maiden names) and aliases (Ishida 2011, Rogers 2018). Because of this, person names may often be transcribed in different ways, in particular if different scripts are involved. As they are also regularly not unique, the enrichment of person data is key for any systematic understanding of the meaning of these data.

Indeed, a particular problem is homonyms. While in taxonomy they can be often addressed with higher taxonomies and in geography with numeric georeferencing or geographic hierarchies (continent, country, state, etc.), for person names various methods of disambiguation may need to be applied with each use case. To avoid this, unique identifiers for persons should be used instead of their names. Such identifiers are minted by various resources for various purposes. This includes the Virtual International Authority File (VIAF), intended for librarians, and the International Standard Name Identifier (ISNI) standard, intended for identifying contributors to creative works. Other initiatives include the Open Researcher and Contributor ID (ORCID) for researchers and Wikidata items for any person with sufficient notability. ORCIDs are notable in that they are intended to be minted by the individuals themselves. Wikidata is unique for its openness, allowing easy addition and editing of records. Many other resources mint identifiers for persons, often for people active in a certain specific field (such as Entomologists of the World, Zoobank or the International Plant Names Index).

One part of the work done in the pilot described in section 2.2.3. (automated collector matching) included an analysis of properties of Wikidata items for persons. For Wikidata person records which could be connected to botanical specimens, hundreds of unique properties with “ID” in their label could be noted, linking to numerous different person authority resources. This proliferation of authority resources, depending on their sources and interlinkages, promises a greater coverage of individual persons. It may also prove an obstacle, as mappings between the different resources may need to be made and it can become difficult to identify problems of synonyms and homonyms across dozens or more of different platforms. There may also be technical obstacles, as not all resources will follow the same data model, architecture and rules for dealing with merges or deletions. Most importantly, not all identifiers may be (intended to be) stable.

3.3. What is our context?

We have established the nature of the information we’re looking for and what we want to achieve. However, not all specimens are born digitally in the same manner. Some have no digital counterpart yet. Others have existed digitally for decades already. Some may have considerable amounts of data available and achieve high MIDS (Minimum Information about a Digital Specimen) levels, whereas others may be stub records with little more than an identifier linking it to the physical entity (Hardisty et al. 2020b). This context is an important aspect to consider when planning enrichment activities.

When a physical specimen is digitized, it may be easy to incorporate an enrichment step as part of the digitization workflow. Of course, the nature of this step depends on the workflow. If specimens are imaged but little other information (like the three properties in focus) is processed and added, enrichment will have a lower priority than the digitization of said information. If this digitization is implemented, for instance through text transcription by volunteers, experts or algorithms, an enrichment step can be incorporated in this particular workflow. Human transcribers can make use of checklists, based on the authorities detailed in the previous section, to add names of people, locations and taxa. Human input can offer additional expertise of the collection, the country or the taxon group.

Algorithms can be constrained to only allow the output of checked values or their output can be processed in turn for enrichment activities (e.g. by lookup, by clustering, followed by language processing, followed by matching to authority identifiers). Incorporating enrichment in the digitization process has the advantage of avoiding double-work, as there are no separate data transcription and data enrichment processes. A disadvantage is that it is more difficult to infer information from other specimens, which may still be in the queue for digitization or reside in other physical collections. This may still lead to double-work, for instance when duplicate specimens were sent to different collections such as is common in botany. It may also lead to enrichments of lower quality, as some specimen properties can only be accurately inferred from information associated with other specimens.

It's a different situation when information has already been digitized. Performing enrichment post-digitization brings the advantage that any information already available can be utilized. Clustered lists of names, locations and taxa can be processed, reducing overall workload. A popular program to process such lists by supervised matching is OpenRefine (Delpeuch 2019). Automated methods are also possible, as shown by the matching scripts described in sections 2.2.3. and 2.6. However, much will depend on the state of the data already available. Data may have been digitized over long time periods, using different methodologies and standards. Such datasets may be poorly interoperable and a lot of information difficult to use, or of dubious quality. For example, specimen collection dates and field numbers may seem promising for inferring collector identities or collection locations. However, they may prove more trouble than they're worth if a lot of dates and numbers have been digitized differently, incompletely or even incorrectly.

Another important consideration for the enrichment context is that enriching specimen data requires certain resources to do so. In the case of human-in-the-loop workflows, it requires that such humans are available, sufficiently trained and up to the task at the scale that is required (which may be massive, with the number of tasks running in six or more digits). Automation is useful to address issues of scale, but it requires the expertise to set up the workflow as well as the hardware to run it, then process and store the outcomes. For collections lacking these resources, automated enrichment can be set up at a further downstream level. Digital records for specimens are planned to enter the European Collection Objects Index (ECOI) that is currently planned as part of the DiSSCo infrastructure (Hardisty et al. 2020a). Services operating at the level of this index could provide a level of automated enrichment. This has already been demonstrated with the DiSSCo Digitiser Software working from Darwin Core Archive files. Such enrichment can also benefit from a greater data availability at this level, allowing more and stronger inferences of information. On the other hand, it may be difficult to keep track of error propagation at this level and quality policies need to be implemented to deal with how these enriched data are to be validated.

3.4. How to enrich?

The effective enrichment, that is attaching standardized values or persistent, unique identifiers to digital specimen records, can be done in multiple ways. The workflow can be mostly manual, where a person goes through each value of the property to enrich one by one. To achieve this, they can consult a preset list of resources where the property is identified, such as Wikidata or ORCID for person data or the Catalogue of Life for taxon data. These resources can be consulted manually or be implemented in an interface for easy lookups, possibly making use of APIs supported by these resources. The latter also raises the possibility of suggestive matches for the manual enricher to validate. This sort of workflow incorporates human input in a workflow that is otherwise quite automated.

Alternatively, human enrichers can also employ a heuristic approach to find any resource where the property in question is identified or additional information can be obtained to refine its identity, for example the use of maps to interpret locality descriptions or skimming through biographies to confirm a person match. Such approaches can be very intensive and time-consuming, but they also pose a challenge that may be appealing for some people and incentivize them to provide enrichment services on a voluntary basis (i.e. crowdsourcing). Bionomia (see section 2.5) is an example of how this approach can work.

Human enrichers can go out of their way and seek out additional information if they are unsure or if they fail to find an initial match. This is much less evident when employing algorithms to enrich properties. Any information to be used needs to be established beforehand and often needs to be cleaned or converted into a machine-readable form. Hence, while these methods can process large numbers of records quite quickly, they require more elaborate pre-processing and tend to work better if the data to enrich and to inform on the enriching is already rather tidy and interoperable. In particular the data to inform on the enrichment are important, such as date ranges or associated specimens, as they play a critical role in disambiguation and validation.

The **first question** to ask is what data to make use of. Oftentimes, a frequency table is created which lists all the unique values for a certain type of information (e.g. all unique strings of person names) and their occurrence frequency. This allows a manual enricher to process the most common values first. This approach saves a lot of time and duplicate work, but it has the downside that homonyms may not be caught (i.e. identical strings referring to different entities). A method to avoid this problem is a more sophisticated approach than a simple frequency table, where the strings are disambiguated based on other pieces of information. This can include dates, regions or taxonomic groups. A clustering approach like this also has the potential of avoiding double-work in case many variants for a single person's name are available. It does require more expertise for the preliminary data analysis and may need input from experts familiar with the collection.

The **second question** is which authority resources to consult. This may not have as much impact as would be expected, because many resources readily incorporate links to others as part of the feature profile they provide. If so, as soon as a link is made with a URI for a record in one resource, this URI can be resolved to retrieve URIs for records in other resources. Resolving these URIs may bring up even more identifiers, as we travel along the knowledge graph, allowing data to be harmonized between different collection sources as long as a link was made in both sources with at least one of these identifiers. An important assumption in this regard is that all these links are accurate and correct. Accuracy may be problematic as not all resources have similar granularity.

A few questions will be important when choosing a resource:

- How likely is it that the feature we are looking for is present in this resource?

- Can the resource be consulted very easily, preferably through an API? Are there restrictions that may hinder the enrichment process? Is the resource free?
- Does the resource offer links to other resources?
- How regularly is the resource being maintained (kept up to date)? Can it be expected to stay alive and stable in the distant future?

A key problem will be features absent in the resource. This is less of a problem for open resources such as Wikidata, which allow addition and editing of records by the general public. However, Wikidata is intended as a secondary resource, so a primary resource will still be needed. Wikidata can still be a good option if a primary resource can be found, but is not suspected to remain stable. It is also a good gateway to locate primary resources, given its powerful APIs and open data model.

When the dataset to enrich has been assembled and the target authority resources fixed, a workflow needs to be agreed upon for linking records to unique identifiers for their different properties. As stated before, a human may process the records one by one or an algorithm can process them instead. Either way, a validation and/or disambiguation step will be needed somewhere in the workflow, as different problems can come up:

3.4.1. No match can be found

This means that either the feature is not present or it failed to be found. A validation step is needed to confirm absence before an alternative enrichment is sought. If the resource is open, like Wikidata or Geonames, adding new records is fairly easy. For semi-open resources like ORCID, a certain procedure may need to be followed (i.e. tracking down the person in question and requesting they create an ORCID for themselves), which will be more time-consuming. If the process of adding missing records becomes too time-consuming or unreliable, alternative resources will need to be consulted.

Another possibility for ‘no match’ is that the feature is not identifiable through the information provided. Depending on the method used to consult the authority resources (see 3.5), this may return “no match” or multiple matches that require disambiguation. If a feature seems to be unidentifiable, it is recommended to look for expertise in the subject at hand, as there may be many reasons for this. Possible solutions include translations from unsupported languages, investigations in gazetteers of old place names, consulting taxonomic experts or seeking additional sources related to the specimen, such as field notebooks or similar specimens. Of course, if still applicable, people who have been involved in the specimen’s collection and curation can be consulted.

3.4.2. A single match is found

This is, of course, the desired outcome, but validation may still be needed to weed out a false positive. A false positive is more likely if the consulted resource contains more features which are definitely incorrect. This is for instance the case when consulting a generic database, where it is known that the majority of records will not fit the profile. It may not be known specifically which records fit and which do not. ORCID is an example of this, as the vast majority of ORCIDs identify people who have not been involved at all in collecting natural history specimens. Of course, once a significant amount of enrichment has been performed in the field of Natural History collections, it will become easier to use these enriched properties to separate likely and less likely records in these resources.

In general, validation will be assumed to be part of the manual enricher's workflow. The enricher can have a quick look at additional data on the subject and decide whether the match makes sense or not. It is possible for additional validation by other individuals, in particular if manual enrichment is performed by multiple people who may slightly differ in their heuristic approach or interpretation of the enrichment protocol. However, this will double the impact of what is typically already the most time- and resource-consuming part of the workflow.

Validation is much more important for the outcome of automated matching methods, in particular if they are suspected or known to be prone to false positives. This limits their usefulness, but it is still easier and faster to check the outcome of a matching process than it is to find all the links one by one yourself. Confidence scores can be useful to streamline the validation process. They can be used as cutoff points for acceptable matches, but also to indicate suspect matches which may very well be false. Manual enrichers may also flag records of which they are unsure the enrichment is accurate.

3.4.3. Multiple matches are found

When multiple matches are found, some disambiguation will be required. Confidence scores can be used for this. It is also possible that these multiple matches are correct and the authority resource contains duplicate records. Duplicate records may have been merged after being identified as duplicates, but their URIs can still resolve and may still be found in other sources (such as Wikidata).

If multiple matches are found and they each identify different features, steps need to be taken to find the correct feature. This can be done by looking at additional data or incorporating a post-hoc validation step in the matching algorithm. It is also possible to pick the most confident match and omit all others.

It may not be possible for some features to be disambiguated. If a specimen collector is only identified through a common name as Jim or James Smith with no further information, disambiguation is not possible. Similarly, common locality names with no further information (e.g. Springfield) might occur. Disambiguating such records mostly lies in the hands of experts, although clustering algorithms may also elucidate patterns that are otherwise not apparent to a human enricher, in particular if they can train on big datasets of specimen data. The more specimens the algorithm can process, the more likely related ones are included, which can inform on the specimen's as such unclear properties.

It is difficult to proscribe an exact protocol as to how disambiguation of natural history specimens is supposed to be performed. This also makes it rather difficult to automatize for general use. The main problem is that different data elements that might support disambiguation can be present at both sides of the enrichment process (source data and target resource). A lack of interoperability too remains a common problem (Dillen et al. 2019). For persons, the following elements might be used:

- Known aliases of this person. This includes common ways they have signed their work or have been described, including abbreviations.
- Period the person was active versus period the specimen was acted upon by the person. This may be further refined with the period the person was known to be in a region where the specimen was present at the time.
- Regions where the person is known to have worked.
- Known collaborators or institutional affiliations.
- Syntax or structure of the collection number assigned to a specimen.

Whether any of these elements are taken into account during the enrichment process depends on their availability and their estimated quality. Database managers or collection curators may have good insights into which elements are reliable and which are not. Digitization protocols may be informative as well. Many of these elements will also require pre-processing to allow them to be easily consulted, in particular if these are to be incorporated automatically in the enrichment workflow. This pre-processing may constitute a massive data cleaning operation, with considerable implications for the running costs of the enrichment process.

For taxon names, the inherent hierarchy can be taken into account. Higher taxon names may be documented (e.g. family name listed on a label) or may be inferred using very basic species recognition. Oftentimes, homonyms occur across kingdoms and not within. In this case, the nature of the specimen (e.g. different preparation methods for plants, fungi or insects) or metadata of the collection it’s housed in will easily allow disambiguation. Synonyms will be more problematic to deal with, but taxon names are still easier to process with validation tools than person names, as a global standard exists for them. Some potential tools have been described in section 3.1. Exact taxon identification is still the work of experts. Promising approaches for automated species recognition have been shown, but lots of work is yet to be done (Carranza-Rojas et al. 2017, Little et al. 2020). However this lies outside the scope of semantic enrichment.

For georeferencing, the process can be tremendously difficult, either for a human or an automated approach. Location descriptions may mention local place names at widely contrasting granularity (e.g. a large province or forest versus a small village or a notable landmark). They may also constitute elaborate descriptions, even directions, of where the specimen was gathered. Understanding them may not be straightforward and require heuristic approaches, such as consulting intuitively associated resources, trial-and-error queries or consulting specific experts (e.g. locals from the suspected area). Place names may have numerous homonyms (e.g. Wikipedia’s list of popular place names) and without additional information these may be impossible to disambiguate. Old synonyms can also occur, as well as different languages. Properties that can help in refining georeferencing or disambiguating potential locations are:

- Collector numbers. Oftentimes, these may have a chronological order and hence location info can be inferred if it is known for other specimens, collected by the same people and which have a closely related collector number.
- Dates. There is a limit to how far a collector can travel in a certain time period, so specimens collected around the same time by the same people will constrain the possible area.
- Regions where the collectors are known to have collected, or where this species is known to occur.
- Habitat descriptions and other ecological elements. Taxonomy may also aid the georeferencing process, as many species have geographical restrictions in where they occur.

3.5. The matching process

Numerous algorithms have been developed for disambiguation of features, in various fields of research. The problem of disambiguation is regularly historical in nature, as the modern digital age has facilitated the use of unique identifiers to identify features rather than common names. Hence, oftentimes unique identifiers for properties can be associated during digitization itself and no further disambiguation is required other than at the level of the authority source. For instance, an iNaturalist user submitting an observation in the app on their phone can easily automatically enrich it with the exact coordinates where it occurred. The observation will be tied to the id of the user’s account, which can in turn be associated with an external authority such as ORCID. Finally, an automatically suggested or expertly proposed identification with a taxon name can be accepted by the user. Regardless of whether the identification is accurate or correct, its meaning will be enhanced as the taxon ids in the iNaturalist taxonomic backbone refer back to authority sources such as the Catalogue of Life.

This example can be extended to the gatherings of specimens, which are simply particular cases of observations. The systems keeping track of specimens accessioned into a collection can offer similar services as iNaturalist to automatically enrich the identity of the collector and a taxonomic identification for the specimen. This will require consulting external authority sources such as ORCID or Catalogue of Life whenever data is added, or at least maintaining up-to-date local copies. For location information, handheld GPS devices have become ubiquitous, even just the lower precision tools found in common smartphones. Even if remote fieldwork complicates the georeferencing of specimens, various resources are now available to post-hoc enrich specimens, such as Google Maps or OpenStreetMap.

Still, an identifier is only as strong as the information connected to it. As more data become available, the odds of false positives may increase. People may have multiple ORCIDs and other identifiers, taxonomic relations may change and these changes need to be properly propagated. Geographic features may require more complex methods of enrichment, where different methods may not be easily compatible (for instance grid-based ecological observations versus point-radius single observations).

Examples of other fields where disambiguation has been investigated include author names of (scientific) publications, names of inventors listed on patents held by official institutions, geographic place names and tags in web search indexes (Garcia et al 2009, Hussain and Asghar 2017, Deyun and Kazuyuki 2018). Different issues may not have (fully) compatible solutions, as the nature of the available data and its meaningfulness may vary, as well as the frequency of problems of homonyms versus synonyms.

Numerous approaches to disambiguation have been developed and used. Hussain and Asghar (2017) proposed a classification system to describe these different approaches. Machine Learning methods are designed to learn how to interpret the data that they are supposed to process. They can achieve this by learning how to associate the input data with the correctly disambiguated output, processing training data in which the right connections have already been made. This approach is called supervised learning, as it requires a training dataset in which correct disambiguations have already been made.

Another Machine Learning approach consists of learning the common patterns in the input data. These patterns can then be used to automatically determine likely clusters or relationships in the dataset to process. The latter approach does not require that training datasets are available, i.e. data which have already been correctly processed so that both the input and the correctly disambiguated output are available. This is called unsupervised learning. Such an approach still requires that the output is interpreted to identify the meaning of the identified clusters and relationships. Approaches that combine both techniques are called semi-supervised learning and can, for instance, consist of a supervised learning algorithm with only a small training dataset processing the results from an unsupervised algorithm. This is an attractive approach as training datasets can be difficult and time-consuming to produce.

Hussain and Asghar (2017) also identify non Machine Learning approaches. Experts in the subject may apply a rules-based workflow to disambiguate name strings based on common patterns, syntax and other bits of knowledge and expertise. Such a workflow will often be heuristic, but may achieve very good results at smaller scales as it is most adept at incorporating various bits of relevant information that are already known. Another approach is the creation of graphs to illustrate the different clusters and relationships to process, visualizing the problem of disambiguation. Graphs are best suited when groups of features need to be disambiguated, such as multiple persons associated with the same specimen or multiple taxa on the same sheet.

The matching process may also be fully manual or incorporate a human element. This includes the use of crowdsourcing. Citizen science platforms have been developed for various bulky tasks, which would be time-consuming for field expert scientists to plough through. A common task is the transcription of label text from a digital image of a specimen, through crowdsourcing platforms such as Digivol, DoeDat or Die Herbonauten. Enrichment may be incorporated as part of this process, for instance by providing checklists sourced from an authority resources, for properties such as specimen collector or taxon name. Gazetteers or georeferencing tools may be implemented as well.

3.6. Publication

After the enrichment process has been successfully undertaken and validated, data will need to be processed so that they can be made available. The most commonly used data standards for specimen data publication is Darwin Core (Wieczorek et al. 2012). Darwin Core is a relatively flat standard (i.e. most relations between properties are one-to-one), extended for specimen data from the Dublin Core properties to describe resources. Specimen records published under the Darwin Core Occurrence standard can have up to one value for each property supported by the standard. Delimitation of multiple values for a single property can be done on an *ad hoc* basis with custom delimiters like pipes (|), as the standard is loose in its enforcement of vocabularies or ontologies for its properties. However, a more consistent way of implementing one-to-many relationships is the use of extensions, which exist for concepts such as multiple taxonomic identifications, multiple images and multiple measurements made of a single occurrence. Extensions consist of additional flat tables linked to the central occurrence table, which can list multiple values for a single property of a single occurrence. However, no further relationships between properties are supported in this “star-schema” structure (Wieczorek et al. 2012).

A few problems arise when trying to fit semantically enriched data to the Darwin Core format. Enrichment will mostly consist of one or more identifiers for a certain value of a property (e.g. an ORCID and a Wikidata ID for a collector of a botanical specimen). While a property exists for this collector, i.e. `dwc:recordedBy`, this is most commonly used for the name(s) of the collector(s), not for identifiers such as an ORCID. No other property currently exists in Darwin Core to represent the collector of a natural history specimen. Sometimes, identifiers are used in the `dwc:recordedBy` field, but this causes interoperability conflicts with the more common use of this field (i.e. person names) and does not follow the recommended Darwin Core best practice. Identifiers could also be concatenated with the verbatim name of the person in question, but this requires a consistent method for concatenation and for later splitting of the string to unambiguously retrieve the identifier. Issues that may arise include:

- Encoding problems: similar characters that are actually different, e.g. | (U+007C) and ∣ (U+2223).
- Delimitation conflicts: e.g. the delimiter occurs as part of the string as well, which requires consistent escaping of the delimiter.
- Software conflicts: Some scripts and algorithms may stumble over unexpected delimiters, in particular if they are commonly used for other purposes such as the pipe (|) as an OR operator.

An alternative is on trial by GBIF, after a release in June 2020: proposed new Darwin Core properties were minted under a GBIF namespace, called `gbif:recordedByID` and `gbif:identifiedByID`. These terms were added to the Occurrence Core, so that in addition to a name string in `dwc:recordedBy`, identifiers could be submitted using these properties. One of the many use cases this enables is for users to find occurrences recorded by a single person through their ORCID, rather than having to search all possible aliases of a name string.

A downside is that only one identifier can be submitted, or concatenation with custom delimiters (commonly |) has to be employed again. A Darwin Core extension for identification of persons is currently under development (see section 2.3). This extension will support listing multiple identifiers for multiple people associated with an occurrence record, including various roles such as “collected” and “identified”. This way, multiple identifiers can be listed as separate values, name strings can be published alongside identifier URIs and groups of people with an identical role (e.g. collector teams) can be listed separately.

An alternative to the use of an extension is the `dwciri` namespace, which was conceived as a parallel to Darwin Core properties (Baskauf and Sachs 2018). The idea was to have a separate namespace for identifiers and a namespace for textual values. This works well if data are structured in a non-tabular way, like when Darwin Core properties are published in an RDF format (Darwin Core and RDF/OWL Task Groups 2015). Multiple identifiers can be listed as additional `dwciri` values of the same property, or `sameAs` properties (as in the `owl` namespace) can be used. Practical examples of this can be found in the Botany Pilot (see section 2.1). A downside of this approach is the inconsistency between extant Darwin Core terms that represent identifiers (e.g. `dwc:institutionID`, `dwc:scientificNameID`, `dwc:higherGeographyID` and the new `gbif:recordedByID`) and their `dwciri` counterparts. There currently is also no straightforward and consistent way to represent `dwciri` values in a typical (and most popular) Darwin Core tabular format. XML or JSON representations can make successful use of the `dwciri` namespace, such as in the CETAF Specimen Preview Profile or in the openDS standard that is currently in development for the European Collection Object Index (ECOI).

3.7. Data maintenance

Enriched data can get published in many ways and through different pipelines. Many (big) collection-holding institutions openly publish their data in institutional or national data portals. Most also publish to GBIF, through an IPT or Biocase server similar in hosting to the data portals. The European Collection Objects Index, a planned product of DiSSCo, will also make specimen data available on the web in the form of digital objects, which will hold information or links to information related to the specimen.

Current practice today is that all these web-available data are a product of a locally maintained data store (sometimes called a Collection Management System, CMS), where the most accurate and extensively annotated digital representation of the physical specimen is being maintained. This allows the institution responsible for the curation of the physical object to maintain control over its digital counterpart. However, this may not always be the case in the future as distribution and decentralisation of enrichment and digital curation work into the expert community becomes more common.

Some CMSs may offer APIs that allow direct requests to the data store, reducing the impact of synchronization issues between different nodes in the publication process. Depending on various factors, including data quality, institutional policies, technical resources and data standard comprehensiveness, the mismatch between what is openly available on the web and what is available in the local CMS may be large or small.

Various problems with these CMS systems have been described in detail elsewhere (Dillen et al. 2019). They are often old, the result of in-house, bespoke development and may not support various functionalities important for modern data management. The support for enriched data elements may be poor or nonexistent. In principle, it should be easy to add database fields and tables to accommodate the enriched versions of record properties, such as identifiers and standardized values that fit well-accepted data standards, vocabularies or ontologies. In practice, the local data model may not support these additions so easily on every level, for instance because the features in question are already covered by the data model in a manner not easily congruent with the enriched properties. One example is an internal taxonomic backbone that is incomplete and/or out of date, making it difficult to readily connect the identifiers from authority sources to the database records. Others are the modeling of multiple entities linked to a single record (e.g. collector teams) or the definition of geographic features.

In addition, support for these systems may be poor or expensive, meaning that the implementation of support for enrichment may not be affordable for the institution, moneywise or in terms of allocating sufficient staff time to the problem. As a result, the outcomes of enrichment may be (temporarily) stored elsewhere, complicating data management and in particular propagation of data updates (further enrichment, validation, cleaning) and version control. As a common example, specimen data currently get validated in many ways by aggregators such as GBIF. The enrichment tool Bionomia (see section 2.5) uses these data aggregated by GBIF to connect specimen occurrences to the people collecting them. However, none of these validations and enrichment processes get fed back into the CMS.

There is another problem with the propagation of data updates (‘roundtripping’): Who maintains the authority of making and/or accepting updates, and how? Traditionally, this was managed or delegated by the institution housing the physical specimens and curating them in local systems. However, once data on these specimens gets distributed openly, feedback from the community or from automated systems of enhancement (often called ‘bots’) will need to be processed and in some way authorized. In particular with automated systems, the number of annotations will become insurmountable for institution staff to manage. An example of an open system making use of community curation is Wikidata. It tackles the validation problem by basing its data model fundamentally on a versioning structure, where any change is logged as a new version and rollbacks are relatively straightforward to implement. Git is another model for information management (typically software code), where a history of changes and their provenance is at the core of the model.

Keeping a record of provenance and changes of records is also in the scope of the DiSSCo infrastructure. As described above, problems with synchronizing these provenance records with local CMS will again arise. Enrichment, in particular, poses a challenge, as data may very well become more dynamic once enrichment processes are introduced. This includes additions of other links than those covered in this report, like genetic sequences or digital derivatives such as Optical Character Recognition (OCR) outputs. But even identifiers for persons or taxa may prove dynamic: errors may be flagged in any part of the graph and will need to be propagated: e.g. a person may be erroneously associated with a specimen or a person record in Wikidata may turn out to be a combination of two different people. Identifiers may also simply turn out not to be as stable or persistent as expected.

4. Conclusions

Semantic enrichment of natural history specimen data has increased in prominence over the last few years and data standards have also begun to accommodate this novel representation of specimen properties. Various tools are now available to facilitate enrichment, from elaborate human interfaces for manual enrichment to algorithms for automated processing. Nevertheless, a universal solution is not possible. Principally, any enrichment workflow needs to commence by answering the questions of (1) what the data currently look like, (2) which specific enrichment targets are desired and (3) what resources can be used to perform and support the effective enrichment process. The answers to these questions will inform which approaches are most attractive.

For example, smaller collections with only very rudimentary digitized specimens may not have the resources for elaborate enrichment workflows. They may ask a staff member to process the most common values or most prominent specimens and enrich them manually in the course of a few days at most, only employing a very minimal protocol. While this may seem trivial, significant portions of the collection may already be addressed this way. Depending on a collection’s history and how well it has been (digitally) curated, there may be a lot of low-hanging fruit.

Larger collections regularly achieving MIDS levels of 2 to 3 may have the means to set up automated workflows to enrich their data, either taking some of the known specifics of their collection into account during pre-processing of the data, building it into the scripts or applying this experience during post-hoc validation by dedicated staff or volunteer enthusiasts on crowdsourcing platforms. Depending on the geographic scope and the history of a specific collection, different resources may be considered as optimal targets for the enrichment process. For instance, a collection may have many specimens collected by relatively well-known scientists who are well-represented across many authority resources, or it may have many collectors who are poorly known in the Anglosaxon world and hence poorly represented in these general resources. The taxonomic coverage is similarly important, and may be even more complicated if specimens in the collection have mostly been classified using older taxonomic backbones.

In general, enrichment should be incorporated into the digitization process as much as possible. This is evidently not possible for specimens already digitized and may complicate existing digitization pipelines, but it saves a lot of double-work and hence resources, while also rendering the digitized product much more meaningful and hence useful for scientific research. In particular already elaborate digitization activities, such as transcription through crowdsourcing, by dedicated staff or as part of the specimen gathering process, should ensure their protocols support enrichment as much as possible. The most common obstacle to this integration is the absence of records in the resources consulted. To mitigate this, open resources such as Wikidata or Geonames should take preference to others, and the digitization pipeline should include a record addition step for these missing records.

In the last years, we've seen tremendous advances in big data science. Complex clustering and enrichment algorithms will also be adapted and applied to natural history specimens, allowing researchers to make links at unprecedented scale and potentially absolve individual collections from implementing their own *ad hoc* enrichment workflows. Automated species recognition or collector identification by handwriting recognition are examples of some outcomes of these algorithms. Given the computing resources, the data availability and the skills required to implement these algorithms, this is done best at a level of considerable data aggregation, such as in GBIF or in the DiSSCo ECOI. Even if these approaches start to deliver, they will not eliminate the importance of local enrichment activities. While elaborate automated enrichment workflows will turn out to have less added value when performed locally, local enrichment activities may still allow curators and other local experts to provide their own experience and knowledge to the Biodiversity Knowledge Graph. And, of course, enrichment done as soon as possible after the specimen gathering event is the best assurance that data have not been misinterpreted or corrupted.

From a technical point of view, large-scale enrichment activities come with some challenges. While most of these are well-understood by IT professionals and addressed extensively in other big data activities, a few (potential) obstacles may prove daunting, in particular when they involve human elements. One such obstacle is the question of enrichment stability. The use of external identifiers to disambiguate specimen properties implies a dependency on the stability of these identifiers. If they break or drift, it may not be straightforward to retrieve the information identified or even the disambiguation it implied. It is impossible to predict which resources will remain stable in the future and which will not, although some are less likely to fail than others, in particular influenced by the level of their use and the (financial) dependencies resting on their stability (e.g. ORCID, which is commonly used by many of the world's scientific publishers). It is also resource-intensive to regularly validate the stability of all identifiers, never mind their identification. A portfolio approach seems the most prudent to mitigate this risk, where multiple identifiers from different resources should be used to identify a single property.

Another technical obstacle is the streamlined flow of data. If enrichment is done at a central, aggregated level, these enhanced data may not seamlessly flow back into the local systems where the specimens are physically curated. Depending on the quality control of central enrichment and other annotation activities, local curators may also wish to validate amendments to the data on their specimens, which may constitute a massive bottleneck. In general, support for enrichment should be considered as a key requirement in future development on collection management systems and institutional or national data portal development. To avoid the validation bottleneck, enrichment should be supervised at the aggregated level, so its results can at least flow back seamlessly to local systems.

Finally, while data standards have adapted to support enrichment in multiple ways, it is clear that the introduction of this enhanced data presentation strains the classic tabular model such as is used in Darwin Core archives. While simple tabular representations still have clear advantages for human interpretation, markup document (XML or JSON) or even relational representations (see section 2.7) will need to be more commonly adopted if enriched properties are to be generally used. An RDF/XML representation such as described by the CETAF Specimen Preview profile or a JSON representation such as proposed for openDS are good examples of how this could work.

5. References

Adamowicz SJ (2015) International Barcode of Life: Evolution of a global research community. Genome. 58(5): 151-162. doi:10.1139/gen-2015-0094

APA (2020) Publication Manual of the American Psychological Association, Seventh Edition (2020)

Baskauf SJ and Sachs J (2018) Darwin Core as a Vocabulary for Expressing Biodiversity Data as RDF. In: Application of Semantic Technology in Biodiversity Science (ed. A. Thessen) IOS Press. doi:10.3233/978-1-61499-854-9-15

Belhajjame K et al. (2013) PROV-O: The PROV Ontology. Lebo T, Sahoo S and McGuinness D (eds.). W3C. https://www.w3.org/TR/prov-o/

Besnard G, Gaudeul M, Lavergne S, Muller S, Rouhan G, Sukhorukov AP, Vanderpoorten A and Jabbour F (2018) Herbarium-based science in the twenty-first century, Botany Letters, 165:3-4, 323-327, doi:10.1080/23818107.2018.1482783

Boyle B, Hopkins N, Lu Z et al. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14, 16. doi:10.1186/1471-2105-14-16

Carranza-Rojas J, Goeau H, Bonnet P et al. (2017) Going deeper in the automated identification of Herbarium specimens. BMC Evol Biol 17, 181. doi:10.1186/s12862-017-1014-z

Chapman AD and Wieczorek JR (2020) Georeferencing Best Practices. Copenhagen: GBIF Secretariat. doi:10.15468/doc-gg7h-s853

Darwin Core and RDF/OWL Task Groups (2015) Darwin Core RDF guide. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/guides/rdf/

DataCite Metadata Working Group (2018) DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.2. DataCite e.V. doi:10.5438/bmjt-bx77.

Delpeuch A (2019) A survey of OpenRefine reconciliation services. arXiv preprint arXiv:1906.08092.

Deyun Y and Kazuyuki M (2018) Inventor Name Disambiguation with Gradient Boosting Decision Tree and Inventor Mobility in China (1985-2016), Discussion papers 18018, Research Institute of Economy, Trade and Industry (RIETI).

Dillen M, Groom Q and Hardisty A (2019) Interoperability of Collection Management Systems. ICEDIG Deliverable 4.4. doi:10.5281/zenodo.3361598

Garcia A, Szomszor M, Alani H and Corcho O (2009) Preliminary results in tag disambiguation using DBpedia. In: The Fifth International Conference on Knowledge Capture (K-Cap'09) - First InternationalWorkshop on Collective Knowledge Capturing and Representation (CKCaR'09), 1 Sep 2009, Redondo Beach,California, USA.

Groom Q, Dillen M, Hardy H, Phillips S, Willemse L and Wu Z (2019) Improved standardization of transcribed digital specimen data, Database 2019: baz129, doi:10.1093/database/baz129

Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE and Haston E (2020) People are essential to linking biodiversity data. Database 2020: in press. doi:10.1093/database/baaa072

Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith VS and Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects, Database 2017, bax003, doi:10.1093/database/bax003

Hardisty A, Saarenmaa H, Casino A, Dillen M, Gödderz K, Groom Q, Hardy H, Koureas D, Nieva de la Hidalga A, Paul DL, Runnel V, Vermeersch X, van Walsum M and Willemse L (2020a) Conceptual design blueprint for the DiSSCo digitization infrastructure - DELIVERABLE D8.1. Research Ideas and Outcomes 6: e54280. doi:10.3897/rio.6.e54280

Hardisty A, Addink W, Dillen M, Groom Q, Haston E et al. (2020b) (Draft) Minimum Information about a Digital Specimen (MIDS) v0.12, 03 November 2020. https://github.com/tdwg/mids

Hilse H-W and Kothe J (2006) Implementing persistent identifiers: overview of concepts, guidelines and recommendations. London / Amsterdam: Consortium of European Libraries and European Commission on Preservation and Access, 2006. ISBN 90-6984-508-3.

Hugo W, Le Franc Y, Coen G, Parland-von Essen J, & Bonino L (2020) D2.5 FAIR Semantics Recommendations Second Iteration (Version 1.0). Zenodo. doi:10.5281/zenodo.4314321

Hussain I and Asghar S (2017) A survey of author name disambiguation techniques: 2010–2016. The Knowledge Engineering Review, 32, E22. doi:10.1017/S0269888917000182

ICZN (1999) International Code of Zoological Nomenclature. Fourth edition. International Trust for Zoological Nomenclature, London, xxix + 306 pp.

Ishida R (2011) Personal names around the world. W3C. Accessed on 2020-12-15 https://www.w3.org/International/questions/qa-personal-names

Jin J and Yang J (2020) BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. Global Ecology and Conservation 21: e00852. doi:10.1016/j.gecco.2019.e00852

Little DP, Tulig M, Tan KC, Liu Y, Belongie S, Kaeser-Chen C et al. (2020) An algorithm competition for automatic species identification from herbarium specimens. Applications in Plant Sciences, 8(6), e11365. doi:10.1002/aps3.11365

Marcer A, Haston E, Groom Q et al. (2020) Quality issues in georeferencing: From physical collections to digital data repositories for ecological research. Divers Distrib. 2020; 00: 1– 4. doi:10.1111/ddi.13208

Meise Botanic Garden (2020) Meise Botanic Garden Herbarium (BR). Version 1.16. Botanic Garden Meise. Occurrence dataset doi:10.15468/wrthhx accessed via GBIF.org on 2021-01-05.

MLA Handbook (2016) Eighth edition. New York: The Modern Language Association of America, 2016. Print.

Rogers T (2018) Falsehoods Programmers Believe About Names – With Examples. Accessed on 2020-12-15 https://shinesolutions.com/2018/01/08/falsehoods-programmers-believe-about-names-with-examples/

Roskov Y, Ower G, Orrell T, Nicolson D, Bailly N, Kirk PM, Bourgoin T, DeWalt RE, Decock W, van Nieukerken EJ and Penev L (eds.) (2020) Species 2000 & ITIS Catalogue of Life, 2020-12-01. Digital resource at www.catalogueoflife.org. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-8858.

Thessen AE, Woodburn M, Koureas D, Paul D, Conlon M, Shorthouse DP and Ramdeen S (2019) Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. Data Science Journal, 18(1), p.54. doi:10.5334/dsj-2019-054

Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Kusber W-H, Li D-Z, Marhold K, May TW, McNeill J, Monro AM, Prado J, Price MJ and Smith GF (eds.) (2018) International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Glashütten: Koeltz Botanical Books. doi:10.12705/Code.2018

University of Chicago Press (2017). Chicago manual of style (17th ed.).

WFO (2019) World Flora Online. Version 2019.05. Published on the Internet; http://www.worldfloraonline.org. Accessed on 2020-12-22

Wieczorek J, Bloom D, Guralnick R et al. (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS One. 2012;7(1):e29715. doi:10.1371/journal.pone.0029715

Wilkinson MD, Dumontier M, Aalbersberg IJ et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, vol. 160018. doi:10.1038/sdata.2016.18

6. Appendix

6.1. Agent attribution extension JSON files

As the GBIF test environment is not intended for permanent linking, the referenced JSON responses including agent attribution extension data are represented here.
Example 1

```
{
  "acceptedNameUsage": {},
  "basisOfRecord": "PreservedSpecimen",
  "catalogNumber": "BR0000025667882V",
  "class": "Magnoliopsida",
  "coordinateUncertaintyInMeters": {},
  "country": "Canada",
  "countryCode": "CA",
  "datasetID": {},
  "datasetName": "Meise Botanic Garden Herbarium",
  "day": "27",
  "decimalLatitude": {},
  "decimalLongitude": {},
  "eventDate": "2017-09-27",
  "extensions": {
    "dwc:Identification": [],
    "https://tdwg.github.io/attribution/people/dwc/AgentActions": [
      {
        "action": "collected",
        "alternateName": "Groom Q. & Engledow H.",
        "displayOrder": "2",
        "endedAtTime": "2017-09-27",
        "name": "Henry Engledow",
        "startedAtTime": "2017-09-27",
        "type": "Person",
        "verbatimName": "Quentin Groom & Henry Engledow"
      },
      {
        "action": "collected",
        "agentIdentifierType": "ORCID",
        "alternateName": "Groom Q. & Engledow H.",
        "displayOrder": "1",
        "endedAtTime": "2017-09-27",
        "identifier": "https://orcid.org/0000-0002-0596-5376",
        "name": "Quentin Groom",
        "startedAtTime": "2017-09-27",
        "type": "Person",
        "verbatimName": "Quentin Groom & Henry Engledow"
      },
      {
        "action": "collected",
        "agentIdentifierType": "wikidata",
        "alternateName": "Groom Q. & Engledow H.",
        "displayOrder": "1",
        "endedAtTime": "2017-09-27",
        "identifier": "http://www.wikidata.org/entity/Q28913658",
        "name": "Quentin Groom",
        "startedAtTime": "2017-09-27",
        "type": "Person",
        "verbatimName": "Quentin Groom & Henry Engledow"
      }
    ]
  },
  "family": "Oxalidaceae",
  "genus": "Oxalis",
  "habitat": "Lawn in front of offices",
  "id": "https://www.botanicalcollections.be/specimen/BR0000025667882V",
  "informationWithheld": "decimalLatitude, decimalLongitude, locality",
  "institutionCode": "BR",
  "institutionID": "http://biocol.org/urn:lsid:biocol.org:col:15605",
  "kingdom": "Plantae",
  "license": "http://creativecommons.org/licenses/by/4.0/",
  "locality": {},
  "locationRemarks": {},
  "modified": "2019-02-01",
  "month": "09",
  "nomenclaturalCode": "ICBN",
  "occurrenceID":
"https://www.botanicalcollections.be/specimen/BR0000025667882V",
  "order": "Oxalidales",
  "phylum": "Tracheophyta",
  "preparations": "HerbariumSheet",
  "recordNumber": "1707",
  "recordedBy": "Groom Q. & Engledow H.",
  "rightsHolder": "Meise Botanic Garden",
  "scientificName": "Oxalis dillenii Jacq.",
  "specificEpithet": "dillenii",
  "taxonomicStatus": "accepted name",
  "type": "PhysicalObject",
  "typeStatus": {},
  "verbatimEventDate": {},
  "year": "2017"
}
```

Example 2

```
{
  "acceptedNameUsage": {},
  "basisOfRecord": "PreservedSpecimen",
  "catalogNumber": "BR0000016884175",
  "class": "Magnoliopsida",
  "coordinateUncertaintyInMeters": {},
  "country": "Ethiopia",
  "countryCode": "ET",
  "datasetID": {},
  "datasetName": "Meise Botanic Garden Herbarium",
  "day": "12",
  "decimalLatitude": {},
  "decimalLongitude": {},
  "eventDate": "1973-01-12",
  "extensions": {
    "dwc:Identification": [],
    "https://tdwg.github.io/attribution/people/dwc/AgentActions": [
      {
        "action": "collected",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "2",
        "endedAtTime": "1973-01-12",
        "name": "A. Getachew",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "3",
        "endedAtTime": "1973-01-12",
        "name": "F. Rasmussen",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "4",
        "endedAtTime": "1973-01-12",
        "name": "Kaj Vollesen",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "agentIdentifierType": "HUH GUID",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "1",
        "endedAtTime": "1973-01-12",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/5dce3a0a-aef2-4788-8830-88474
570932a",
        "name": "Ib Friis",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "agentIdentifierType": "ORCID",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "1",
        "endedAtTime": "1973-01-12",
        "identifier": "https://orcid.org/0000-0002-2438-1528",
        "name": "Ib Friis",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "agentIdentifierType": "VIAF",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "1",
        "endedAtTime": "1973-01-12",
        "identifier": "https://viaf.org/viaf/161032823/",
        "name": "Ib Friis",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "agentIdentifierType": "VIAF",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "1",
        "endedAtTime": "1973-01-12",
        "identifier": "https://viaf.org/viaf/326154387298130970004/",
        "name": "Ib Friis",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      },
      {
        "action": "collected",
        "agentIdentifierType": "wikidata",
        "alternateName": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
        "displayOrder": "1",
        "endedAtTime": "1973-01-12",
        "identifier": "http://www.wikidata.org/entity/Q3043068",
        "name": "Ib Friis",
        "startedAtTime": "1973-01-12",
        "type": "Person",
        "verbatimName": "I. Friis, Getachew A., F. Rasmussen & K. Vollesen"
      }
    ]
  },
  "family": "Malvaceae",
  "genus": "Abutilon",
  "habitat": {},
  "id": "https://www.botanicalcollections.be/specimen/BR0000016884175",
  "informationWithheld": "decimalLatitude, decimalLongitude, locality,
locationRemarks",
  "institutionCode": "BR",
  "institutionID": "http://biocol.org/urn:lsid:biocol.org:col:15605",
  "kingdom": "Plantae",
  "license": "http://creativecommons.org/licenses/by/4.0/",
  "locality": {},
  "locationRemarks": {},
  "modified": "2017-12-31",
  "month": "01",
  "nomenclaturalCode": "ICBN",
  "occurrenceID": "https://www.botanicalcollections.be/specimen/BR0000016884175",
  "order": "Malvales",
  "phylum": "Tracheophyta",
  "preparations": "HerbariumSheet",
  "recordNumber": "2219",
  "recordedBy": "Friis I., Getachew A., Rasmussen F. & Vollesen K.",
  "rightsHolder": "Meise Botanic Garden",
  "scientificName": "Abutilon longicuspe var. cecilii (N.E.Br.) Verdc.",
  "specificEpithet": "longicuspe",
  "taxonomicStatus": "unchecked name",
  "type": "PhysicalObject",
  "typeStatus": {},
  "verbatimEventDate": {},
  "year": "1973"
}
```

Example 3

```
{
  "acceptedNameUsage": {},
  "basisOfRecord": "PreservedSpecimen",
  "catalogNumber": "BR5020011434859",
  "class": "Agaricomycetes",
  "coordinateUncertaintyInMeters": {},
  "country": "Congo, Democratic Republic of the",
  "countryCode": "CD",
  "datasetID": {},
  "datasetName": "Meise Botanic Garden Herbarium",
  "day": {},
  "decimalLatitude": {},
  "decimalLongitude": {},
  "eventDate": "1930-08",
  "extensions": {
    "dwc:Identification": [],
    "https://tdwg.github.io/attribution/people/dwc/AgentActions": [
      {
        "action": "identified",
        "agentIdentifierType": "BHL",
        "endedAtTime": "1963",
        "identifier": "https://www.biodiversitylibrary.org/creator/179458",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "BHL",
        "endedAtTime": "1964",
        "identifier": "https://www.biodiversitylibrary.org/creator/179458",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "BHL",
        "endedAtTime": "1965-03",
      "identifier": "https://www.biodiversitylibrary.org/creator/179458",
      "name": "Rolf Singer",
      "startedAtTime": "1965-03",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "BHL",
      "endedAtTime": "1976",
      "identifier": "https://www.biodiversitylibrary.org/creator/179458",
      "name": "Rolf Singer",
      "startedAtTime": "1976",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "BHL",
      "endedAtTime": "1963",
      "identifier": "https://www.biodiversitylibrary.org/creator/2529",
      "name": "Rolf Singer",
      "startedAtTime": "1963",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "BHL",
      "endedAtTime": "1964",
      "identifier": "https://www.biodiversitylibrary.org/creator/2529",
      "name": "Rolf Singer",
      "startedAtTime": "1964",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "BHL",
      "endedAtTime": "1965-03",
      "identifier": "https://www.biodiversitylibrary.org/creator/2529",
      "name": "Rolf Singer",
      "startedAtTime": "1965-03",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "BHL",
      "endedAtTime": "1976",
      "identifier": "https://www.biodiversitylibrary.org/creator/2529",
      "name": "Rolf Singer",
      "startedAtTime": "1976",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "HUH",
      "endedAtTime": "1963",
      "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=19208
",
      "name": "Rolf Singer",
      "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH",
        "endedAtTime": "1964",
        "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=19208
",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH",
        "endedAtTime": "1965-03",
        "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=19208
",
        "name": "Rolf Singer",
        "startedAtTime": "1965-03",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH",
        "endedAtTime": "1976",
        "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=19208
",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH",
        "endedAtTime": "1930-08",
        "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=62330
",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "HUH",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier":
"https://kiki.huh.harvard.edu/databases/botanist_search.php?mode=details&id=62330
",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1930-08",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/2acba475-4ffd-487a-9eac-55fe4
258f8b8",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "HUH GUID",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/2acba475-4ffd-487a-9eac-55fe4
258f8b8",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1930-08",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/99684b11-4399-41b0-9b74-0f6ba
665bd9a",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "HUH GUID",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/99684b11-4399-41b0-9b74-0f6ba
665bd9a",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1963",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/bfcddcf1-7be0-48f2-a93f-88381
5688079",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1964",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/bfcddcf1-7be0-48f2-a93f-88381
5688079",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1965-03",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/bfcddcf1-7be0-48f2-a93f-88381
5688079",
        "name": "Rolf Singer",
        "startedAtTime": "1965-03",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "HUH GUID",
        "endedAtTime": "1976",
        "identifier":
"http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/bfcddcf1-7be0-48f2-a93f-88381
5688079",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "IPNI",
        "endedAtTime": "1930-08",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=10020-1",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "IPNI",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=10020-1",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "IPNI",
        "endedAtTime": "1963",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=9693-1",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "IPNI",
        "endedAtTime": "1964",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=9693-1",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "IPNI",
        "endedAtTime": "1965-03",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=9693-1",
        "name": "Rolf Singer",
        "startedAtTime": "1965-03",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "IPNI",
        "endedAtTime": "1976",
        "identifier": "https://www.ipni.org/ipni/idAuthorSearch.do?id=9693-1",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "ISNI",
        "endedAtTime": "1963",
        "identifier": "http://isni.org/isni/0000000355357632",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "ISNI",
        "endedAtTime": "1964",
        "identifier": "http://isni.org/isni/0000000355357632",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "ISNI",
        "endedAtTime": "1965-03",
        "identifier": "http://isni.org/isni/0000000355357632",
        "name": "Rolf Singer",
        "startedAtTime": "1965-03",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "ISNI",
        "endedAtTime": "1976",
        "identifier": "http://isni.org/isni/0000000355357632",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1963",
        "identifier": "http://viaf.org/viaf/92597180",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1964",
        "identifier": "http://viaf.org/viaf/92597180",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1965-03",
        "identifier": "http://viaf.org/viaf/92597180",
        "name": "Rolf Singer",
        "startedAtTime": "1965-03",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1976",
        "identifier": "http://viaf.org/viaf/92597180",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1930-08",
        "identifier": "https://viaf.org/viaf/188552669/",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "VIAF",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier": "https://viaf.org/viaf/188552669/",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "VIAF",
        "endedAtTime": "1930-08",
        "identifier": "https://viaf.org/viaf/311294421/",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "VIAF",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier": "https://viaf.org/viaf/311294421/",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "wikidata",
        "endedAtTime": "1930-08",
        "identifier": "http://www.wikidata.org/entity/Q10349681",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "collected",
        "agentIdentifierType": "wikidata",
        "alternateName": "Staner P.",
        "endedAtTime": "1930-08",
        "identifier": "http://www.wikidata.org/entity/Q10349681",
        "name": "Pierre Joseph Staner",
        "startedAtTime": "1930-08",
        "type": "Person",
        "verbatimName": "Staner P."
      },
      {
        "action": "identified",
        "agentIdentifierType": "wikidata",
        "endedAtTime": "1963",
        "identifier": "http://www.wikidata.org/entity/Q64091",
        "name": "Rolf Singer",
        "startedAtTime": "1963",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "wikidata",
        "endedAtTime": "1964",
        "identifier": "http://www.wikidata.org/entity/Q64091",
        "name": "Rolf Singer",
        "startedAtTime": "1964",
        "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "wikidata",
        "endedAtTime": "1965-03",
      "identifier": "http://www.wikidata.org/entity/Q64091",
      "name": "Rolf Singer",
      "startedAtTime": "1965-03",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "wikidata",
      "endedAtTime": "1976",
      "identifier": "http://www.wikidata.org/entity/Q64091",
      "name": "Rolf Singer",
      "startedAtTime": "1976",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "wikispecies",
      "endedAtTime": "1930-08",
      "identifier": "https://species.wikimedia.org/wiki/Pierre_Staner",
      "name": "Pierre Joseph Staner",
      "startedAtTime": "1930-08",
      "type": "Person",
      "verbatimName": "Staner P."
    },
    {
      "action": "collected",
      "agentIdentifierType": "wikispecies",
      "alternateName": "Staner P.",
      "endedAtTime": "1930-08",
      "identifier": "https://species.wikimedia.org/wiki/Pierre_Staner",
      "name": "Pierre Joseph Staner",
      "startedAtTime": "1930-08",
      "type": "Person",
      "verbatimName": "Staner P."
    },
    {
      "action": "identified",
      "agentIdentifierType": "wikispecies",
      "endedAtTime": "1963",
      "identifier": "https://species.wikimedia.org/wiki/Rolf_Singer",
      "name": "Rolf Singer",
      "startedAtTime": "1963",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "wikispecies",
      "endedAtTime": "1964",
      "identifier": "https://species.wikimedia.org/wiki/Rolf_Singer",
      "name": "Rolf Singer",
      "startedAtTime": "1964",
      "type": "Person",
      "verbatimName": "Singer R."
    },
    {
      "action": "identified",
      "agentIdentifierType": "wikispecies",
      "endedAtTime": "1965-03",
      "identifier": "https://species.wikimedia.org/wiki/Rolf_Singer",
      "name": "Rolf Singer",
      "startedAtTime": "1965-03",
      "type": "Person",
        "verbatimName": "Singer R."
      },
      {
        "action": "identified",
        "agentIdentifierType": "wikispecies",
        "endedAtTime": "1976",
        "identifier": "https://species.wikimedia.org/wiki/Rolf_Singer",
        "name": "Rolf Singer",
        "startedAtTime": "1976",
        "type": "Person",
        "verbatimName": "Singer R."
      }
    ]
  },
  "family": "Marasmiaceae",
  "genus": "Marasmius",
  "habitat": {},
  "id": "https://www.botanicalcollections.be/specimen/BR5020011434859",
  "informationWithheld": "decimalLatitude, decimalLongitude, locality,
locationRemarks",
  "institutionCode": "BR",
  "institutionID": "http://biocol.org/urn:lsid:biocol.org:col:15605",
  "kingdom": "Fungi",
  "license": "http://creativecommons.org/licenses/by/4.0/",
  "locality": {},
  "locationRemarks": {},
  "modified": "2007-05-16",
  "month": "08",
  "nomenclaturalCode": "ICBN",
  "occurrenceID": "https://www.botanicalcollections.be/specimen/BR5020011434859",
  "order": "Agaricales",
  "phylum": "Basidiomycota",
  "preparations": "HerbariumSheet",
  "recordNumber": "325",
  "recordedBy": "Staner P.",
  "rightsHolder": "Meise Botanic Garden",
  "scientificName": "Marasmius crinisequi F.Muell.",
  "specificEpithet": "crinisequi",
  "taxonomicStatus": "accepted name",
  "type": "PhysicalObject",
  "typeStatus": "Type of Marasmius crinisequi var. monocotyledonum Singer",
  "verbatimEventDate": "19300800",
  "year": "1930"
}
```