The design and prototype of a workflow integrating Wikidata into validation and linking

Milestone MS32

30 June 2023

Authors

Mathias Dillen[1], Andreas Plank[2]

[1]: Meise Botanic Garden, Meise, Belgium
[2]: Botanical Garden and Botanical Museum, Berlin, Germany

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

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

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Milestone title:	The design and prototype of a workflow integrating Wikidata into validation and linking
Milestone n°:	MS32
Means of verification:	Report
WP responsible:	WP7
Lead beneficiary:	MeiseBG
Citation:	Dillen, D. & Plank, A. (2023). *The design and prototype of a workflow integrating Wikidata into validation and linking*. Milestone MS32 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of milestone:	Month 25
Actual submission date:	30 June 2023

Table of contents

Preface 4
Summary 5
List of abbreviations 5
1. Roundtripping links 6
    1.1. Bionomia 6
        1.1.1. Introduction 6
        1.1.2. Roundtripping attributions 7
        1.1.3. Mapping 9
        1.1.4. Issues 10
    1.2. Wikidata 11
        1.2.1. Models for specimens 11
        1.2.2. Conclusion 12
2. The current state of Person PIDs 12
    2.1. GBIF 12
    2.2. Botany Pilot 16
    2.3. Bionomia 17
    2.4. Refining and validating matches 17
3. Linking workflows 19
    3.1. Clustering 19
    3.2. Fuzzy string matching 21
        3.2.1. Wikidata subset 22
        3.2.2. Name strings 22
        3.2.3. String parsing 23
        3.2.4. Matching 23
4. Acknowledgements 23
5. References 23

Preface

Wikidata is an open community-curated linked database with no explicit focus in terms of content, hosted by the Wikimedia foundation. As such, it is an appealing resource to broker unambiguous and persistent identifiers for concepts that are not specific to the field of biodiversity, such as people or geographic entities. People obviously play an important role in biodiversity research, but as data they are also key in linking related data elements together, in space and time, or in terms of provenance, such as taxa described in literature, which itself cites physical specimens curated by collection-holding institutions. People that are tied to those specimens, such as those who collected them or refined ('identified') the taxonomic identity of these collected specimens are one of the common denominators throughout these series of different data elements and may thus be of great help in effectively connecting them together.

Much work has been initiated throughout the past years to disambiguate people in biodiversity, by linking them to persistent identifiers, rather than referencing them by various different incarnations of their name(s). Global initiatives such as ORCID have helped address the issue primarily in literature, whereas specifically in the context of biodiversity, projects such as the CETAF-based Botany Pilot and the citizen science platform Bionomia have kickstarted and promoted the linking process. GBIF has implemented some basic support for linking as well, which by now has also been ratified into the Darwin Core standard, but work is still underway in TDWG to extend this formally.

Summary

In this task, the aim is to develop a workflow that should facilitate the linking process of collector name strings to PIDs for those collectors. Such a workflow should help scale up the number of links being made, make the process more efficient and should take advantage as much as possible of existing work and infrastructures, so as not to reinvent the wheel. As such, the work can be roughly split into a few subtasks:

- Make existing linking workflows more easily implementable in other contexts and by other infrastructures. This includes finding ways for such workflows to produce links that can easily be published, i.e. in a standardised format compatible with existing infrastructure. The suitability of different infrastructures for making established links available should also be assessed.
- Establish, document and improve the comprehensiveness, findability and interoperability of the content in PID-minting resources, in particular Wikidata as it can be edited openly.
- Refine the decision making process of establishing links, by implementing and improving the methods that can be used to validate potential links.

In this document, the focus lies on linking people. We will propose a workflow to 'roundtrip' links established through the Bionomia platform back to the collections holding the attributed specimens, as well as making them available for use by other BiCIKL infrastructures. We will also refine existing automated linking workflows and pilot the new functionalities on the (botanical) collections of the task partners. These refinements will be influenced by an assessment of the current state of Wikidata, investigated through shape expressions constructed from commonly used queries and from Wikidata records which have been linked in previous efforts such as the Botany Pilot, Bionomia and published specimen data to GBIF.

List of abbreviations

ABCD	Access to Biological Collection Data
ALA	Atlas of Living Australia
BGBM	Botanical Garden and Botanical Museum Berlin
BHL	Biodiversity Heritage Library
CETAF	Consortium of European Taxonomic Facilities
CMS	Collection Management System
DiSSCo	Distributed System of Scientific Collections
DwC	Darwin Core
EU	European Union
GBIF	Global Biodiversity Information Facility
IPNI	International Plant Names Index
MeiseBG	Meise Botanic Garden
ORCID	Open Researcher and Contributor IDPub

openDS	Open Digital Specimens
PID	Persistent Identifier
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
TDWG	Biodiversity Information Standards
VIAF	Virtual International Authority File

1. Roundtripping links

Roundtripping is a term that has many different meanings depending on the context. In the context of data science, it has been popularized (see e.g. this and this) as a term to refer to the process of data being taken from a resource to be enriched, validated or improved. A successful roundtrip means that the data subsequently makes it back into the original resource without any ambiguity or synchronization problems.

Successful roundtripping is a question of interoperable data models and standards, as well as protocols for conflicts that may arise. The main complications in the context of natural history specimens are the poor interoperability between different data managing infrastructures, in particular the diverse set of local institutional databases (Dillen et al. 2019), the lack of persistent identifiers for specimens (Agosti et al. 2022) and gaps in the existing standards.

1.1. Bionomia

1.1.1. Introduction

Bionomia is a web platform developed and maintained by David P. Shorthouse. It was launched in 2018 and has since then facilitated the attribution of millions of specimens published to GBIF to the people who collected or determined them, identified through either their ORCID or a Wikidata item record. ORCID is the preferred authority resource to reference, as the IDs are minted by the people they identify themselves. As this does not work for (long) deceased people, Wikidata is leveraged as an alternative.

Every two weeks, Bionomia ingests specimen occurrence data from GBIF and the name strings attached to those records, through the Darwin Core properties `recordedBy` and `identifiedBy`. Since their introduction in 2020, the platform also ingests PIDs already published to GBIF through the Darwin core properties `recordedByID` and `identifiedByID` and validates them if they're Wikidata or ORCIDs.

In addition to the GBIF data, Bionomia also keeps caches of subsets of Wikidata and ORCID. The subset of Wikidata is defined by the presence of a death date and the presence of at least one external link to an authority source that is considered sufficiently connected to the field of specimen collecting and determining. The cache of ORCID is based on keywords.

Table 1: *Properties and keywords used to find relevant records in Wikidata and ORCID.*

Wikidata property	Property ID	ORCID Keywords
IPNI	P586	*taxonomy*
Harvard Index of Botanists	P6264	*taxonomist*
Entomologists of the World	P5370	*mycology*
ZooBank Author ID	P2006	*zoology*
BHL Creator ID	P4081	*entomology*
Stuttgart Database of Scientific Illustrators ID	P2349	*botany*
Zürich Herbaria collector ID	P10168	*systematics*
		phylogenetics
		biodiversity

The strings gathered from GBIF are then processed to try and capture different name elements from them and identify multiple names of different people (i.e. teams). The code that performs this is also available as a standalone ruby gem and Bionomia offers an API that implements it. The processed strings from GBIF are then matched to the Wikidata and ORCID caches. Volunteers can validate suggested matches and make new ones through the platform. Various query services and statistics are incorporated to facilitate this validation process. The platform also hosts some metadata for people, in particular their affiliation, mostly taken from ORCID and from GBIF occurrence records.

1.1.2. Roundtripping attributions

Attributions made through Bionomia can be downloaded in multiple ways. At the level of a GBIF dataset, the easiest way to retrieve them is as a Frictionless Data package based on the Darwin Core archive standard, but with additional metadata and more relations between elements. The package contains the following files:

- `datapackage.json`: Metadata for each other file, including timestamp for when the package was last produced
- `users.csv`: all attributed people with their parsed names and date metadata
- `occurrences.csv`: attributed occurrence records from GBIF with a selection of Darwin Core properties
- `problem_collector_dates.csv`: occurrence records of which the eventDate clashes with the lifespan of the attributed collector
- `attributions.csv`: all attributions made/validated for occurrence records, including attribution metadata
- `missing_attributions.csv`: all attributions that were not sourced from GBIF itself (from recordedByID/identifiedByID)
- `unresolved_users.csv`: attributions sourced from GBIF that Bionomia could not validate through ORCID/Wikidata. Includes values for recordedByID/identifiedByID that are not Wikidata or ORCID URIs (e.g. VIAF).

These attributions can be retrieved and roundtripped back to the source, i.e. GBIF or the systems that provide the data to GBIF. This roundtripping is still not very common, given the small number of datasets that provide recordedByIDs or identifiedByIDs to GBIF (215 out of the 22,821 processed by Bionomia, as of 2023-06-28). To increase the visibility of the Bionomia attributions, browser extensions have been produced for Firefox and Chrome that show on GBIF pages the ORCIDs or Wikidata IDs that have been attached to recordedBy/identifiedBy values in the Bionomia platform. However, this is simply a client-side page modification by the browser, retrieving any attributions as JSON-LD from the Bionomia API based on the GBIF occurrence record or dataset keys. The attributions will not be available through the GBIF API.

Roundtripping may not happen for a variety of reasons, such as a lack of resources at the data provider level or poor interoperability between different data models, for instance when trying to import attributions back into the local CMS. Attributions are not very well supported in the Darwin Core standard. An extension for Darwin Core archives to support detailed and one-to-many attributions of agent actions to occurrences was in development the past few years and made it to a test implementation within GBIF's User Acceptance Testing environment (Dillen et al. 2021). Development has since stalled and no further support is available other than the additions of recordedByID/identifiedByID or leveraging generic extensions such as dynamicProperties or resourceRelationship.

GBIF has since been working on a new model, the GBIF Unified Model, for the data it aggregates, considerably more relational and flexible than Darwin Core, supporting more complex data types. This includes support for agent attributions related to any other data class, as well as PIDs for those agents. Similarly, DiSSCo has been further developing its openDS data model, which features agent as one of the elements. Current implementations do not support PIDs yet for these agents, although PIDs can be added as annotations, such as in this example.

```
"data": [
  {
    "id": "20.5000.1025/KGN-9JV-ZZS",
    "type": "Annotation",
    "attributes": {
      "id": "20.5000.1025/KGN-9JV-ZZS",
      "version": 1,
      "type": "Annotation",
      "motivation": "linking",
      "target": {
        "id": "https://hdl.handle.net/20.5000.1025/6E8-BHL-31V",
        "type": "digital_specimen",
        "indvProp": "ods:collector"
      },
      "body": {
        "type": "ods:collector",
        "value": [
          "https://www.wikidata.org/wiki/Q61390"
        ],
        "description": "wikidata"
      },
```


Fig. 1: *Example of an annotation in the current prototype of the DiSSCo Infrastructure.*

Collection specimen profiles defined on cetafidentifiers.biowikifarm.net/wiki/CSPP allow RDF properties to be harvested as collection data; those RDF profiles have been harvested into the Botany Pilot (services.bgbm.org/botanypilot/) and its RDF triple store database until the end of 2022. As they are gathered from diverse curatorial data, they could reflect the curational processing. But the current harvest state is not up to date and is only launched periodically on an ad hoc basis (github.com/infinite-dao/glean-cetaf-rdfs/). Bulk harvest endpoints are also not included within the CETAF ID specification, so attributions have to be scraped one-by-one for each individual specimen record. By running SPARQL queries on the Botany Pilot triple store, it is possible to retrieve the harvested and processed collection data from each botanical collection.

1.1.3. Mapping

To facilitate roundtripping, the frictionless data produced through Bionomia can be mapped to existing standards, such as the Darwin Core Agent Attribution extension. The main file is `missing_attributions`. This contains the attributions that need to be roundtripped. To map the main export to the Agent Attribution extension, the following fields need to be populated:

- `agentType`: Can be set to a fixed value of "Person".
- `agentIdentifierType`: Either "wikidata" or "orcid", which can easily be inferred from `identifier`.
- `occurrenceID`: Join these from occurrences.csv.
- `identifier`: maps to "recordedBy" or "identifiedBy" in missing_attributions

- `name`: Join these from users.csv. Some names may be missing from that file.
- `alternateName`: NULL
- `verbatimName`: "recordedBy" or "identifiedBy" from occurrences.csv
- `action`: Either "collected" or "identified".
- `role`: NULL
- `displayOrder`: Bionomia does not support an order for multiple agents attached to a single occurrence record, so this cannot be inferred from its data package.
- `identificationID`: Only applicable if the Darwin Core Verifications extension is used, which Bionomia does not seem to ingest
- `startedAtTime`: Either "eventDate" or "dateIdentified" from occurrences.csv
- `endedAtTime`: Either "eventDate" or "dateIdentified" from occurrences.csv

An R script was made that performs this mapping and produces an Agents Attribution extension file for a Darwin Core archive. This script can be extended later to also support mapping to the GBIF Unified Model and to DiSSCo's openDS model, as soon as both are mature enough.

1.1.4. Issues

It is not currently possible for displayOrder to be ascertained without any ambiguity. This occurs because attributions in Bionomia are made at the level of a string attached to a specimen record. If this string encompasses more than one person, multiple attributions of different individuals (and their PIDs) can be tied to the specimen record (and its collector/determiner string). But these attributions do not specify which part of the string they are disambiguating by linking it to a persistent identifier. This is also not a straightforward task to perform, as team strings may be textually structured in different ways.

Connecting attributions of determiner actions to these determinations is also not possible if the determinations were published on GBIF through the Identification History extension to Darwin Core. Bionomia also misses some enriched attributions on GBIF this way, as this extension supports the identifiedByID property. The number of Preserved Specimens in GBIF that make use of this extension is over 20M, so it is possible a significant number of links are missed this way.

The Agents Attribution extension is not implemented and not ratified. A key problem (and why it has stalled) is the definition of what an agent action is. A single agent may play multiple roles, perform multiple actions, at different times, on a specimen, and may be identified by different PIDs. It is not easy to address this problem within the constraints of Darwin Core, as a 'bag of terms' with very limited relationships between properties. It is far more likely that the gap in standards will be addressed when new developments such as the GBIF Unified Model and DiSSCo's openDS reach sufficient Technical Readiness.

Attributions will in practice be linked to specimens in various ways, e.g. by gbifID, but also by occurrenceID, catalogNumber, etc. This complicates a generic workflow and causes links to break. Steps have been taken at GBIF to minimize this. The latest ingests into Bionomia have seen considerable drops in broken gbifIDs (cf. for instance this tweet by the platform). DiSSCo's planned infrastructure for Digital Specimen PIDs will also solve this problem of fuzzy and breaking identifiers (Addink et al. 2023).

1.2. Wikidata

1.2.1. Models for specimens

Specimens themselves are not in the scope of Wikidata, as they are most of the time not sufficiently notable and, most importantly, there are too many of them. Wikidata currently contains a few 100 million items, whereas the estimated number of specimens in the world runs into billions. Taxonomically important specimens may be included, and this has been investigated in the past and has been done to some extent already. The following SPARQL query retrieves all items claimed to be "type specimens" for example, 814 as of 2023-06-21:

```
SELECT ?item ?itemLabel
WHERE
{
  ?item wdt:P31 wd:Q51255340.
  SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en". }
}
```

However, the way taxonomy is modelled in Wikidata is not without problems and this makes incorporating type specimens into the ecosystem a tedious and controversial effort. A way to sidestep this is by populating the type specimen records mostly with external links, rather than internal ones to other Wikidata items and not making use of (potentially flawed) Wikidata properties.

Wikidata can still be extremely useful even if specimens are mostly covered elsewhere and taxonomy is challenging. Specimen-related data, such as geographical locations, people, institutions and their collections, all often have wider applications than collection curation. Hence, they may meet the threshold for inclusion into Wikidata and may, in fact, already be there. The challenge lies in unambiguously retrieving these records in Wikidata, as the generic scope of the system implies a high risk of false positives. Hence the method of querying for a subset of Wikidata, as implemented by Bionomia and others, by requiring the presence of one or more item claims, such as `instance of human` or certain biodiversity-related identifier schemes.

Specimen related people can potentially be retrieved in various ways:

- occupation -> botanist, botanical collector, taxonomist...
- affiliation -> known natural history institutes
    - P485 (archives at) can be used, but there are other ways persons can be linked to an affiliated institution
- authors of taxonomic articles
- **collection items at (P11146)**
- has works in the collection (P6379): but this property is meant to be used for created items such as artworks or books, so it should not be used for our use case.
- has a PID from a biodiversity-related resource, e.g. the resources listed in Table 1

Specimen related people can be tied to one specimen or a group of them through properties like P11146. This is still not used very often and the method to justify this claim varies, that is multiple types of references are possible. It can even be seemingly circular when referencing a Bionomia URI. This is not really the case, as the Bionomia Profile is used as a proxy for the

aggregated specimen data from GBIF. The SPARQL query below retrieves all the references currently used for this property.

```
SELECT DISTINCT ?item ?itemLabel ?affiliation ?affiliationLabel ?ref ?refLabel
WHERE {
  ?item p:P11146 ?statement .
  ?statement ps:P11146 ?affiliation .
  ?statement prov:wasDerivedFrom ?refnode .
  ?refnode   pr:P854 ?ref.
  SERVICE wikibase:label {bd:serviceParam wikibase:language "en".}
}
```

1.2.2. Conclusion

P11146 seems like the best generic recourse for tagging people in Wikidata as collectors of Natural History Specimens. Still, the method of justifying such a claim varies (e.g. a Bionomia profile, a GBIF occurrence record, a GBIF dataset, an ALA search query, a CETAF physical specimen identifier, pdfs/books/articles/web pages... Hence, as part of the workflow to (semi-)automatically enrich specimen data with person identifiers and roundtrip these enrichments back to Wikidata, person records in Wikidata that have been linked to at least one citable specimen, will be enriched using the "collection items at" property. The easiest way to do this is probably in batch through Quick Statements.

2. The current state of Person PIDs

The wider adoption of P11146 in these enrichment efforts will increasingly improve the linking process, as, the more links are made, the larger the group of "likely collectors" becomes. To jump-start this process, we may want to refine the queries currently used to subset this "collector space" from Wikidata, to ensure we do not miss important records by current choices. Additionally, we want to investigate the current state of known collectors currently in Wikidata, so that we know which additional types of data (such as dates, locations, affiliations or external links) we can leverage to more reliably link name strings to Wikidata records, using more than just string matching.

To do this, we will investigate the current state of Person PIDs for specimens in the most prominent resources that currently deal with this: GBIF, the Botany Pilot and Bionomia. The code to perform these analyses can be found here.

2.1. GBIF

Using the asynchronous predicate API, data was acquired from GBIF using the following query.json and a curl request. The query is a variant of the query used by Bionomia to ingest data from GBIF, but the data returned is a Darwin Core Archive rather than an Apache AVRO structure. The same data can be downloaded from doi:10.15468/dl.4whtqj.

```
{
  "creator": "[username]",
  "notificationAddresses": [
    "[mail address]"
```

```
  ],
  "sendNotification": true,
  "format": "DWCA",
  "predicate": {
    "type": "and",
    "predicates": [
      {
        "type": "in",
        "key": "BASIS_OF_RECORD",
        "values": [
        "OCCURRENCE",
        "LIVING_SPECIMEN",
        "FOSSIL_SPECIMEN",
        "PRESERVED_SPECIMEN",
        "MATERIAL_SAMPLE",
        "MATERIAL_CITATION"
      ],
       "matchCase": false
      },
      {
     "type": "or",
     "predicates": [
     {
           "type": "isNotNull",
           "parameter": "RECORDED_BY_ID"
           },
           {"type": "isNotNull",
           "parameter": "IDENTIFIED_BY_ID"
           }
           ]
      }
    ]
  }
}
```

As of 2023-06-27, GBIF has 6.738.518 records with either a `dwc:recordedByID` or `dwc:identifiedByID`, excluding observation records with `dwc:basisOfRecord` set to "HumanObservation" or "MachineObservation" (GBIF.org, 27 June 2023). This also does not include any records for which there may be a `dwc:identifiedByID` value in the Identification History extension, as this field is not easily queryable through the API. Most of these records (> 5M) have `dwc:basisOfRecord` set to "PreservedSpecimen", but there is also a substantial number (> 1M) of records ambiguously labelled as "Occurrence" and more than 400.000 have "MaterialSample".

There are currently 232 datasets providing recordedByID or identifiedByID to GBIF, but about half of the specimens enriched this way originate from three datasets. One of those is Hill and Hobern (2021) and is the source of most records that are listed with Occurrence as `dwc:basisOfRecord`. This dataset indeed does not provide specimen data, but instead lists observations made from light traps. Some of those sampled animals were later accessioned as specimens, but those are not explicitly described in this dataset. If we ignore records listed as "Occurrence", we get the following 12 making up more than 90% of attributed specimens:

Table 2: *Non-observation datasets with most attributions.*

Dataset Key	# Records	%	% cumulativ	Dataset Name
b740eaa0-0679-41dc-acb7-990d562dfa37	1,602,826	28.5%	28.5%	Meise Botanic Garden Herbarium (BR)
e45c7d91-81c6-4455-86e3-2965a5739b1f	795,815	14.2%	42.7%	Vascular Plant Herbarium, Oslo (O) UiO
4ce8e3f9-2546-4af1-b28d-e2eadf05dfd4	572,097	10.1%	52.8%	National Herbarium of Victoria (MEL) AVH data
4bfac3ea-8763-4f4b-a71a-76a6f5f243d3	508,590	9.1%	61.9%	Museum of Comparative Zoology, Harvard University
26f5b360-8770-4d54-9c2d-397798a5e513	386,778	6.9%	68.8%	Entomology, Oslo (O) UiO
ffae417e-b2d8-476c-afe4-8c1093b67071	375,496	6.6%	75.4%	Bacterial members of the Pinus pinaster rhizosphere microbiota in a forest subjected to drought conditions
7948250c-6958-4a29-a670-ed1015b26252	243,946	4.4%	79.8%	Lichen herbarium, Oslo (O) UiO
e4deab67-0998-4140-b573-0ba1f624eb3e	202,874	3.6%	83.4%	Fungarium, Oslo (O) UiO
68a0650f-96ae-499c-8b2a-a4f92c01e4b3	174,819	3.1%	86.5%	Bryophyte Herbarium, Oslo (O) UiO
2b044aa9-1a9a-413e-8b18-ed09da575d3f	95,190	1.7%	88.2%	University of Tartu Natural History Museum and Botanical Garden Zooloogical Collections
a559e942-0fbe-4f09-93ca-28cf244ce2a0	79,674	1.4%	89.6%	Herbarium of University of Coimbra (COI)
30bc94f2-50aa-4688-8e87-a8e11d3d69ff	67,431	1.2%	90.8%	Vascular plant herbarium (KMN) UiA

There is some bias in a classification like this, because some collections may be split into smaller datasets, such as those from UiO, the University of Oslo, whereas others may be published as big blocks, such as the Meise Botanic Garden Herbarium. The five UiO collections in this list, for instance, are good for more than 32% of all records with recordedByID or identifiedByID on GBIF. But these results do indicate that only a few of the larger collections that provide data to GBIF already provide substantial numbers of attributions. So, there is still quite some room for either additional enrichment or roundtripping of existing enrichments that are currently not published (on GBIF). This is particularly relevant in the context of Bionomia, which at this point contains over 30 million attributions (see 2.3), many of which therefore do not originate from or trip back to GBIF.

We can also analyse the types of PIDS used for those two fields. To do this, we identify the most commonly used authority resources and a few other variants (integer numbers and uncategorized "others"). We also split up strings concatenated using semicolons. This produces the following results.

Table 3: *Different types of PIDs found in recordedById/identifiedByID.*

PID type	# Unique PIDs*	# Records
wikidata.org	12,235	1,914,064

orcid.org	10,309	3,033,794
integer	4,138	144,355
viaf.org	909	1,106,364
other	627	40,901
ipni.org	596	108,187
kiki.huh.harvard.edu	257	256,650
scholar.google.com	18	1,533
isni.org	17	1,505
biodiversitylibrary.org	9	2,855
researchgate.net	6	2,742
linkedin.com	4	320
zoobank.org	1	142

Some of these categorised PIDs may not actually be unique, as identifiers are not always provided in the exact same manner. These results have not been cleaned to ensure uniqueness for each PID per resource. For ORCIDs, as an example, the identifier may be provided as the URI, including the orcid.org domain, or as only the 16 digit code. Different aberrations may also occur that should ideally be discouraged, such as adding clarifying strings to the PID (e.g. "ORCID"), padding spaces, omitting the https protocol from the URI or truncation of the identifier. Different variations of this problem occur for any of the other authority resources and complicate the interoperability of the identifiers. Also, the uniqueness is not additive, as the same persons may exist in different authority files.

The bulk of authority resources can be restricted to ORCID, Wikidata itself, VIAF, IPNI and the Harvard Index of Botanists. Knowing that there is a strong overlap between the Harvard Index and IPNI, we will only look at the other four. IDs from these resources were cleaned to extract only the individual ID as expected by the Wikidata property (i.e. NOT the URI). Doing so, we notice a dramatic reduction in uniqueness, e.g. only 5,474 for Wikidata, only 2,246 for ORCID, 909 for VIAF and 596 for IPNI.

To aggregate all Wikidata representations from the GBIF data, we will use properties corresponding to ORCID, VIAF and IPNI ids in Wikidata to find items linked with these ids. This can be done with the following SPARQL query in batch, using the ORCID property P496 as an example:

```
SELECT ?item ?itemLabel WHERE
      {
      ?item wdt:P496 "'0000-0003-3375-7408".
      SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en" }
      }
```

Doing this, we were able to find 822 for ORCID (37%), 610 for VIAF (67%) and 596 for IPNI (100%). Stacking all those Wikidata item IDs for a total of 7,502, of which 7,055 were unique, we can use the Wikimedia `wbgetentities` endpoint to retrieve all claims of those items in batches of 50. We can then combine all those results to construct custom shape expressions as to how common claims turn out to be for this subset of Wikidata items.

Among those 7,055 Wikidata items, we found 1,066 different properties, most (854) of which were IDs in external resources.

2.2. Botany Pilot

Attributions ingested into the Botany Pilot were accessed by SPARQL queries. The following query was used for each institutionID, in batches of 3 per query. cspp stands for the CETAF Specimen Preview Profile, the standard expected from the RDF data ingested from the various CETAF identifier endpoints. The SPARQL endpoint itself is not fully public, but can be accessed upon request to the maintainers of the triple store at BGBM. Results from these queries can be found in the code repository linked earlier.

```
PREFIX dwciri: <http://rs.tdwg.org/dwc/iri/>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT * WHERE {
  GRAPH ?graph {
      ?cspp dcterms:title ?cspp_title; # title is mandatory for
cspp
            dwc:recordedBy ?recordedBy ;
            dwc:institutionID  ?institutionID ;
            dwciri:recordedBy ?recordedBy_IRI ;

      FILTER (
      ( ?institutionID IN (
<https://ror.org/01h1jbk91>,
<https://ror.org/00bv4cx53>,
<https://ror.org/05k35b119>
      )
      )
      )
  }
}
```

In the Botany Pilot triple store, currently attributions are stored for 4.248 different persons (`dwciri:recordedBy` from the CSPP model in the RDF). These attributions originate from 24 different institutions, but more than half of these are delivered through the RDF endpoints of the CETAF identifiers of MeiseBG and BGBM. Through the SPARQL endpoint of this triple store, it is not so easy to quickly determine the number of specimens that have been enriched this way. In the GBIF data export, Meise Botanic Garden had 1.601.355 specimens enriched by a `dwc:recordedByID`, but these data are not perfectly in sync with the Botany Pilot, which probably has a bit less.

Many providers to the Botany Pilot, including BGBM, also provide data to GBIF, but they do this through Biocase and the ABCD standard. Whereas ABCD 3.0 supports person PIDs through the resourceURI property, the GBIF crawler currently does not map this property to its own (Darwin Core-based) model yet. Hence, quite some added value from the Botany Pilot is still to be expected when compared to data from GBIF.

Most of the IDs retrieved from the Botany Pilot were Wikidata IDs (>70%). This is not a very good reflection though, as specimens in the Botany Pilot may also contain additional PIDs using the `owl:sameAs` property. Unfortunately, queries for those additional PIDs timed out,

so we need to find an alternative solution to aggregate them. Retrieving the claims for the Wikidata items lead to a set of 839 different properties.

2.3. Bionomia

Bionomia makes all attributions linked to public profiles available as triples on its downloads page. Using an export retrieved on 2023-06-23, we can note 30,319,617 attributions made for 26,691,799 different occurrences for 11,569 different Wikidata (10,695) or ORCIDs (874). With ORCID, 6,760,277 links have been made, compared to 23,559,340 Wikidata attributions. Given that Wikidata is mainly used to attribute people who have no ORCID, typically because they are (long) deceased, this reflects the considerable historical interest of Bionomia scribes in doing their attributions.

Through the ORCID property P496, Wikidata records corresponding to those ORCIDs were obtained through a similar SPARQL query as shown in the GBIF section, yielding 608 Wikidata items. This leads to a set of unique Wikidata ids of 11,303 items. Aggregating all claims for those items leads to 1,301 different properties.

2.4. Refining and validating matches

Based on these properties, we can try to devise optimal queries to find candidate collector records in Wikidata, and look for reliable properties that could be implemented in a matching workflow to resolve multiple matches for a single name string, or to remove false positives.

The property P11146 we came to recommend in section 1 was not well represented at this point in time, from 6 to 10% of records depending on the source. This property should probably still be included in any subsetting query of Wikidata, as it is the best proxy we have for being a collector, but it will not be a very performant indicator until more enrichment happens and roundtrips back to Wikidata.

We categorize properties in two groups: ones that can improve findability and ones that can improve linking accuracy. Many properties do not fit either group. P21, sex or gender, is commonly present in Wikidata, but is only helpful if we know the gender of the name we want to match. Generic identifiers like ISNI or VIAF ID are also commonly present, but do not help much for either of our use cases. The two groups are listed in the two tables below. Differences between the three sources of attributions were not very big, but sometimes of note. The bias towards botany was clear in all three datasets, but evidently the strongest in the Botany Pilot. The properties in table 5 were generally better represented in the Bionomia dataset, reflecting the added value of the scribes on that platform interacting directly with Wikidata and curating records there as they are attributing. Bionomia IDs (P6944) were not as well represented in the other datasets. This suggests that there is quite some complementarity still between the three resources, even if Bionomia is by far the biggest in number of attributions.

Indeed, of the 14,430 unique Wikidata items listed in either of the three datasets, 5,834 occurred only in Bionomia, 1,648 in GBIF and 1,178 in the Botany Pilot (the rest were at least represented in two of the three datasets).

Table 4: *Properties that occur in > 20% of Wikidata records and that can be used as indicator of affinity for collection work.*

Property	id	GBIF	Bionomia	Botany Pilot
Bionomia ID	P6944	74%	100%	53%
occupation	P106	98%	98%	100%
Harvard Index of Botanists ID	P6264	59%	54%	85%
botanist author abbreviation	P428	60%	43%	86%
IPNI author ID	P586	60%	43%	86%
BHL creator ID	P4081	28%	35%	38%
field of work	P101	22%	23%	30%
Entomologists of the World ID	P5370	8%	16%	11%
Zoobank author ID	P2006	7%	7%	2%
Stuttgart Database of Scientific Illustrators ID	P2349	3%	5%	5%
Zürich Herbaria collector ID	P10168	15%	11%	29%

"botanist author abbreviation" overlaps with IPNI author id. The Bionomia ID operates similarly to P11146 ("collection items at") in that it is essentially a proxy for a direct link to a specimen, i.e. "proof" that the person in Wikidata has at least been linked to a single specimen.
The non-botanical IDs are not very well-represented. Botanical specimens are somewhat over-represented in digital repositories because, as more or less flat herbarium sheets, they are generally easier to digitise than other types. Hence, resources for zoology or entomology are likely to be more important for specimens that have yet to be enriched.

This can also be seen in the non-ID properties here, "occupation" and "field of work". Both of these link to Wikidata items such as Q2374149 ("botanist") or Q441 ("botany"). Items can have multiple "occupation" or "field of work" claims and the possible subjects are several hundred. Some of these are hierarchical, e.g. "botanist" is a subclass of "biologist". But these overarching classes also include many likely false positives, such as "cell biologist" or "computational biologist". Both "botanist" and "zoologist", as well as their many subclasses, should yield mostly relevant results, however.

It becomes clear that a standard such as Latimer Core (Woodburn et al. 2022), which describes collections and their content and is currently in the process of ratification, could be of major help here. If the discipline or taxonomic nature of the specimens to be enriched is known, a different set of SPARQL queries could be utilized depending on the collection metadata.

This means that we will need a generic set of queries for the case when the discipline is not known, and different subsets of this generic set that correspond to different disciplines.

Table 5: *Properties that occur in > 20% of Wikidata records and that can be used to resolve multiple matches or remove false positives.*

Property	id	GBIF	Bionomia	Botany Pilot
given name	P735	86%	92%	91%
date of birth	P569	85%	95%	90%
date of death	P570	70%	94%	74%
country of citizenship	P27	71%	80%	69%
place of birth	P19	45%	62%	59%
family name	P734	54%	67%	56%
languages spoken, written or signed	P1412	33%	37%	49%
place of death	P20	35%	51%	48%
educated at	P69	28%	35%	31%
employer	P108	29%	33%	27%

Here, the most reliable properties are clearly **given name** and **date of birth/death**. Country of citizenship is a tricky one to infer from the source name strings to match, as are other very specific properties such as place of birth/death and "educated at". P1412 could be leveraged if something is known about the language the specimen data was written in, but this is also often not available.

While family name is surprisingly often absent, it can in a relatively reliable manner be inferred from name strings, including Wikidata item labels, and therefore incorporated as an extra filter.

3. Linking workflows

Performing the workflow of linking collector person names to WikiData person names we will use pilot data sets and two already developed name matching workflows: Clustering of names (Klazenga 2020), and Fuzzy Matching (Dillen et al. 2021). We will improve those workflows to ensure they can be applied more easily and widely and to improve their efficacy by taking results from section 2 into account. By making these workflows more generic, they could also be implemented for related types of data, such as collectors or taxonomists referred to in material citations. In Treatmentbank, for instance, almost 300.000 distinct Collector Name strings are present for Material Citations.

3.1. Clustering

The workflow of name match clustering was adapted from Klazenga (2020), which uses basically a partial term frequency comparison (see Borcan (2020) for deeper understanding): it splits author names into ordered name parts, matches those parts for similarity and also it is calculating a name distance: if the name distance is zero, we have a full match, if it is greater than zero we have only partial matching (see detailed workflow scheme below). As a basic matching assistance, at this stage, only the names were matched, but also the collectors lifetime or the collection record dates should be taken into account to refine the matching in time, especially in cases of multiple name matches.

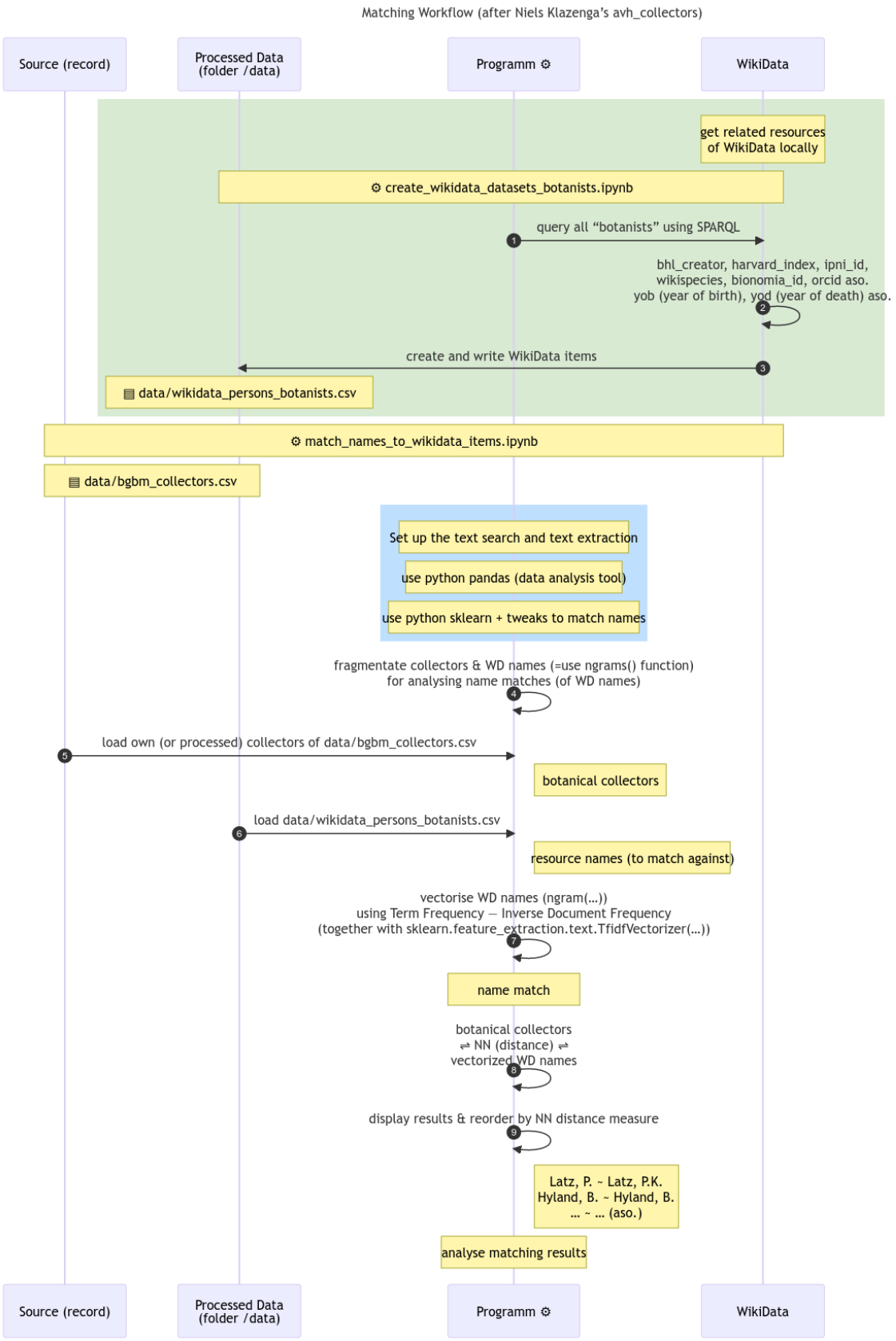

Fig. 2: *Workflow scheme of matching WikiData person names (esp. "botanists") to match with collector names (from BGBM as an example), the code is written in Python (ipynb=Jupyter Notebook): (1) first we query all botanists from WikiData and save them locally; the matching*

(8) uses fragmented names (4...7) and calculates a matching distance (nearest neighbour, NN, (9)); if a NN distance calculates to 0, then it is an exact name string match, otherwise it's partial. Code and idea follows Niels Klazenga's github.com/nielsklazenga/avh-collectors (Australasian Virtual Herbarium, AVH).

The workflow is written mainly in Python (documented as Jupyter notebooks), and the name splitting and parsing uses the Ruby gem package https://libraries.io/rubygems/dwc_agent — the code will be documented at https://github.com/infinite-dao/collector-matching/.

Preliminary results show that name matching in this way is quite good at finding the associated WikiData IDs to enable mapping for your own data. However, for multiple same names but from different people, one has to do a selection by hand, or one could include the floruit time span of the collector to increase the accuracy of correct names.

Preparation of collector name data (input data):

- source of collector names can be any textual data (e.g. tab separated column data)
- splitting long name lists into individual names for better matching (dwc_agent ruby gem)

Addenda:

- include match of date ranges: collecting dates vs. the lifetime of WikiData's person record
- provide access to biological (botanical) collector name data sets, e.g. Botany Pilot (SPARQL), Plazi (https://tb.plazi.org/GgServer/srsStats) to evaluate and refine the matching of own data sets to already reviewed and annotated data sets, and eventually improve data quality of WikiData person records

3.2. Fuzzy string matching

A workflow in R that made use of fuzzy matching to match name strings to a subset of Wikidata items was implemented at Meise Botanic Garden (Dillen et al. 2021). This workflow was designed to make use of properties and local curation protocols in use in Meise's local database system, BG-Base, and specifically its COLLECTORS table. Results of this workflow were incorporated into the general enrichment process at the institution, which also included manual enrichment by collection experts, and the links obtained were made available to the Botany Pilot (through the RDF endpoint on botanicalcollections.be) and to GBIF in the Meise Botanic Garden Herbarium dataset using `dwc:recordedByID`. This workflow was quite successful, as evidenced by the significant contribution of the Meise dataset to the results obtained in section 2, but has not been implemented anywhere else other than through an adapted workflow for the Australasian Virtual Herbarium (see 3.1).

To make the workflow more flexible, less hardcoded for the peculiarities of one collection and more fit for use in other disciplines than Botany, it needs to be adapted to be more generic and effective, and better documented. That will be one of the chief outcomes of this task. In this section of the milestone, we will describe the outline of this adapted workflow, taking the results from section 2 into account.

3.2.1. Wikidata subset

Originally, the matching script made use of six SPARQL queries for certain Wikidata properties, as well as a link to the Wikispecies project. When assessing the importance of each property for the set of enriched items found in section 2 from Bionomia, GBIF and the Botany Pilot, we noted that five of six provided hundreds of items not included in any other query. Only Zoobank was quite sparse in unique contributions.

Table 6: *Number of unique Wikidata Q IDs that were included in the enriched datasets from section 2 and were only present in one of the six Wikidata subsets based on the listed property.*

Property	IDs
Harvard Index of Botanists ID	1037
BHL creator id	304
Wikispecies	261
Entomologists of the World ID	236
IPNI author ID	216
Zoobank ID	9

A relatively large number of items were not found in the union of all those query results (3,452 out of 14,430). Hence, additional queries were set up. Adding queries for "Bionomia ID", "collection items at" and for "occupation" being either "botanist", "zoologist" or any subclass (or further subclass) of those, reduced the number of uncovered items to 619. Almost 25% of those uncovered were due to merges happening in Wikidata since the records were attributed. The remaining 482 may be too sparsely populated to be reliably subsetted this way. It is also possible that these include some incorrect attributions.

3.2.2. Name strings

Name strings can be provided in different ways. Additional metadata, either inferred from specimens linked to the names, or biographical data stored in a local database, such as a Collectors table, can be added and is of great use to solve ambiguous matches and avoid false positives.

As a result, the base input the workflow should be able to take is simply a list of name strings. These strings should be connected to a specimen (or material citation) PID, so the enrichments can actually be roundtripped, but at the base level this is up to the data provider. They can always join enrichments by the name strings back into the database the strings were exported from.

The most widely used data standard for specimen data is Darwin Core. Hence, an import module that interprets Darwin Core properties and extracts both name strings and inferred dates should be included as well.

A third import module will support any Wikidata mapping. That is, in addition to the name strings, biographical data that match in a simple manner to Wikidata properties can be

explicitly listed using their Wikidata property identifier. The most prominent examples will be date of birth/death.

3.2.3. String parsing

The name strings are parsed using the dwc_agent ruby gem into likely name parts (first and last name and middle names) and separated into different team members.

3.2.4. Matching

These parsed name strings are then matched using fuzzy matching to the records aggregated from Wikidata. Any positive matches are kept. This is a different approach than taken initially in Dillen et al. (2021), where a forking set of rules was used to restrict the volume of data on which fuzzy matching had to be applied. This time, we aim to find as many possible candidate matches as possible. Filtering based on different name elements and additional metadata, such as date of birth/death or floruit dates, will be done afterwards on the subset of candidate matches.

The advantage of taking this approach is that the chances of the initial Wikidata SPARQL queries to time out are reduced, as less content needs to be included in them. Also, this opens up the matching process to many more options, as during the second round of selection, any property can be checked now for each of the candidate items. This can be done against a local cache aggregated through `wbgetentities` and parallel to the initial matching process.

The workflow is implemented in R and also makes use of Ruby. It is still a work in progress, but its latest version can be found here.

4. Acknowledgements

We thank Sharif Islam and Annika Hendriksen from Naturalis Biodiversity Center for their helpful feedback.

5. References

Addink W, Islam S, Dillen M, Güntsch A, Theocharides S (2023) Deliverable D7.1 Architecture Design for a pan-European PID system for Digital Specimens. ARPHA Preprints. https://doi.org/10.3897/arphapreprints.e107168

Agosti D, Benichou LL, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8: e97374. https://doi.org/10.3897/rio.8.e97374

Borcan, Marius. 2020. 'TF-IDF Explained And Python Sklearn Implementation'. Medium. 8 June 2020. (What is TF-IDF (Term Frequency — Inverse Document Frequency) and how you can implement it in Python and Scikit-Learn.) https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275.

Dillen, M., Groom, Q., Cubey, R., von Mering, S., & Hardisty, A. (2021). DiSSCo Prepare Deliverable D5.4 A best practice guide for semantic enhancement and improvement of semantic interoperability. https://doi.org/10.34960/AJXS-ZR25

Dillen, M., Groom, Q., & Hardisty, A. (2019). Interoperability of Collection Management Systems. Zenodo. https://doi.org/10.5281/zenodo.3361598

GBIF.org (27 June 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.4whtqj

Groom, Quentin, Christian Bräuchler, Robert Cubey, Mathias Dillen, Pieter Huybrechts, Nicole Kearney, Niels Klazenga, et al. 2022. 'The Disambiguation of People Names in Biological Collections'. Biodiversity Data Journal 10 (October): e86089. https://doi.org/10.3897/BDJ.10.e86089

Hill L, Hobern D (2021). Catches of numerous insect species in Rothamsted 160W light trap at Devonport, Tasmania, 1992-2019. Version 1.0. Atlas of Living Australia. Sampling event dataset https://doi.org/10.5281/zenodo.6820319 accessed via GBIF.org on 2023-06-28.

Klazenga, Niels. (2020). 'AVH Collectors (Australasian Virtual Herbarium)'. Jupyter Notebook. https://github.com/nielsklazenga/avh-collectors

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