Biodiversity Data Journal 10: e86089
doi: 10.3897/BDJ.10.e86089

Methods

The disambiguation of people names in biological collections

Quentin Groom[‡], Christian Bräuchler[§], Robert W. N. Cubey[|], Mathias Dillen[‡], Pieter Huybrechts[‡], Nicole Kearney[¶], Niels Klazenga[#], Siobhan Leachman[¤], Deborah L Paul[«,»], Heather Rogers[˄], Joaquim Santos[˅], David Peter Shorthouse[¦], Alison Vaughan[#], Sabine von Mering[ˀ], Elspeth M Haston[|]

[‡] Meise Botanic Garden, Meise, Belgium
[§] Naturhistorisches Museum Wien, Wien, Austria
[|] Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
[¶] Biodiversity Heritage Library (BHL) Australia, Melbourne, Australia
[#] Royal Botanic Gardens Victoria, Melbourne, Australia
[¤] Independent Researcher, Wellington, New Zealand
[«] University of Illinois, Champaign, United States of America
[»] Florida State University, Tallahassee, United States of America
[˄] McGill University, Montreal, Canada
[˅] Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Coimbra, Portugal
[¦] Agriculture & Agri-Food Canada, Ottawa, Canada
[ˀ] Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

Corresponding author: Quentin Groom (quentin.groom@plantentuinmeise.be)

Academic editor: Jörg Holetschek

Received: 03 May 2022 | Accepted: 26 Aug 2022 | Published: 10 Oct 2022

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

Abstract

Scientific collections have been built by people. For hundreds of years, people have collected, studied, identified, preserved, documented and curated collection specimens. Understanding who those people are is of interest to historians, but much more can be made of these data by other stakeholders once they have been linked to the people's identities and their biographies. Knowing who people are helps us attribute work correctly, validate data and understand the scientific contribution of people and institutions. We can evaluate the work they have done, the interests they have, the places they have worked and what they have created from the specimens they have collected. The problem is that all we know about most of the people associated with collections are their names written on specimens. Disambiguating these people is the challenge that this paper addresses. Disambiguation of people often proves difficult in isolation and can result in staff or researchers independently trying to determine the identity of specific individuals over and over again. By sharing biographical data and building an open, collectively maintained dataset with shared knowledge, expertise and resources, it is possible to collectively deduce the identities of individuals, aggregate biographical information for each person, reduce duplication of effort and share the information locally and globally. The authors of this paper aspire to disambiguate all person names efficiently and fully in all their variations across the entirety of the biological sciences, starting with collections. Towards that vision, this paper has three key aims: to improve the linking, validation, enhancement and valorisation of person-related information within and between collections, databases and publications; to suggest good practice for identifying people involved in biological collections; and to promote coordination amongst all stakeholders, including individuals, natural history collections, institutions, learned societies, government agencies and data aggregators.

Keywords

authority file, attribution, biography, linked open data, Wikidata

Introduction

Biological collections contain a wealth of information on the occurrence of organisms and the people who collected them. As scientific objects of human endeavour, biological specimens rely on people to collect, curate, identify, image and examine them. As records of the existence of taxa at particular points in time and place, they rely on accurate information on the movements and activities of people in order to maximise their fitness-for-purpose. People are central to biological collections and accurate data about the people who collect and care for them is an essential part of the scientific record.

The many and varied traces each of us leave behind in the course of our lives provide clues to our identities, our activities and our movements (Groom et al. 2020). People are identified in writing to give them credit for their work, to provide authority to the statements they make and to link them to their corpus and networks. Publications, field notes, specimen labels, databases, collection registries and logs of research ships all contain the names of people undertaking technical and scientific pursuits. We also leave traces of our activities beyond these pursuits that can help refine the accuracy and repeatability of our science. Birth certificates, death certificates, immigration records, passports, census data, family trees, correspondence, affiliations with institutions and societies and other documents and statements about our lives are all circumscribed within the boundaries of time and space. The names of people are important aids in the discovery of related information. They are also vital entry points into the rich, emotive narratives of human activity, which play a powerful role in relaying the urgency and importance of our scientific work (Weil 2007, Constant and Roberts 2017).

Biographical information about individuals can be used to validate scientific data associated with their work and that of their collaborators. Collecting dates and localities can be verified from personal details such as birth, death and marriage dates, places of residence and employment, personal interests and travels. By using biographical information, redundant digitisation or analytical efforts on related biodiversity data records can be avoided and incorrect or inaccurate information rectified. Missing pieces of data may be inferred, which become easier as more records can be linked through common identifiers. Thus information, such as the key dates in a person's life, the places they visit and the people they know, are important.

Data about people are widely used in historical research, but these data have many other uses in every scientific discipline. Here, we specifically focus on people associated with biological collections. These are mainly people who collect, curate, identify and analyse biological specimens, but they are also people who name new species, study their ecology, undertake a variety of other research and publish their work. Unambiguously identifying people also makes it easier to quantify their contributions, particularly for those, such as taxonomists, for whom citation indices are not necessarily a useful indicator of output, nor quality (Valdecasas 2011). Furthermore, broadening the evaluation of scientific contributions may help increase the diversity and inclusivity of career evaluations (Bhalla 2019, Woolston 2022).

People are typically identified by their names, whether in full or abbreviated. However, people's names are not unique identifiers. It is difficult to know whether a single name string refers to one or more people (Finch 2008, Smalheiser and Torvik 2011) or whether multiple different name strings refer to a single individual. In writing, names are rendered as a sequence of characters with variations in format and script that are often a function of cultural practices. Consequently, one person can be referred to by several different versions of their name, none of which is guaranteed to be unique to that individual. To use biographic data for the purposes outlined above, we need to be able to identify individuals unambiguously. In this paper, we use the term 'disambiguation' to mean the processes by which we resolve any uncertainty in the use of a person's name using any and all corroborating information. The end result of the process is a declaration of the shared identity/ies of a person referred to for any unresolved character string for a person name expressed using the breadth of evidence for the assertion, ideally attached to a persistent identifer.

The process of disambiguation is a particular challenge for natural history collections, where data about individuals – and the specimens collected or annotated by them – are often distributed amongst many different collections, publications and databases. In this paper, we focus on people, though we acknowledge that there are other "agents" that potentially perform roles in collections and on specimens, such as the parent organisation or automated machine processes.

Disambiguation is a process that brings multiple benefits. Clarifying the "who" for specimens extends the number of records useful for research by linking them together; it aids data analysis by helping identify duplicate specimens collected by the one person; it helps resolve past and present collector networks (Groom et al. 2014, Meeus et al. 2021); it supplies collections with information they can use to better resolve their holdings and share with other collections (Groom et al. 2020, Güntsch et al. 2021); it facilitates discovery of a given group of experts through time and space; it provides researchers with additional and direct paths to publishing their collecting and identifying expertise-effort; and, lastly, it demonstrates how the work of each individual contributes to the biodiversity knowledge graph (Page 2016).

The widespread and collaborative nature of biological collections necessitates a shared approach to disambiguation that utilises robust data-sharing mechanisms to avoid the duplication of effort or replication of errors. All people, living or deceased, who have contributed to biological collections ought to have a persistent, unique and freely-sharable identifier. This requires a strategy that reaches well beyond the narrowness of the biological sciences through explicit reliance on other, well-recognised platforms and solutions. Although there is some domain specificity required, the disambiguation of person names in biological data requires a battery of generic tools and services that draw upon multiple lines of evidence, terminating in authoritative, stable and canonical representations of identity. We expect these mechanisms to be equally useful for other domains and to further contribute to the concept of linked open data.

This paper provides particular guidance on the disambiguation of people who, being deceased, are unable to disambiguate their own names. Living people have an incentive to maintain their public biographical data in public resources and have a responsibility to do so if they are employed to generate scientific output. If they are providing citable resources to the scientific record, then they should help maintain the integrity of these data, including their own identity. This does not mean they have to share any personal information, but registering and using an Open Researcher and Contributor ID (ORCID) is a simple step each researcher should take to preserve their own scientific legacy and the legacy of the institution that employs them. Disambiguating one's own identify also avoids generating disambiguation work in the future (Groom et al. 2020).

In this paper, we review and propose best practices for the disambiguation of people in collections. We provide strategies that can be used and considerations for the prioritisation of the work. We outline the biographical resources available for disambiguation, the tools for making the process efficient, the options for unique identifiers and the best practices and recommendations for documenting disambiguation, expressing uncertainty and maintaining these data in databases. We provide examples as case studies and detail the pitfalls that can be encountered. Finally, we suggest some of the uses for these data and consider the possibilities of globally disambiguated collections, including the positive feedback loops of the disambiguation process.

Disambiguation in Society

Ethical and legal considerations

Any dealings with information about people and their activities require careful consideration of the ethical and legal implications of collecting and sharing personal data. Digital technologies have made it easy to gather extensive information on a large number of people. There have always been moral and legal constraints related to the use of data; however, data on people's activities have long been published in biodiversity literature. The power of digital technologies to process these data has prompted governments to enact legislation to formalise the rules regarding the collection and processing of personal data. In Europe, the General Data Protection Regulation (GDPR) is a notable example (Copas 2019, Staunton et al. 2019, European Data Protection Supervisor (EDPS) 2020). However, even in the absence of legal constraints, we have a moral obligation to be sensitive and conscientious when handling personal data. GDPR establishes various legal bases on which data can be kept and processed. One legal basis is for people to give their consent, but another is “legitimate interest” (European Data Protection Supervisor (EDPS) 2020). In the case of those individuals authoring publications and/or collecting and identifying specimens, there is a clear legitimate interest of scientific collections to collate and share relevant data on these people. GDPR also only applies to living people; however, this does not mean there is not a responsibility to be considerate to living relatives of deceased people. GDPR has additional restrictions for “sensitive data”. These are data such as the racial or ethnic origin of a person, their political opinions, religion or philosophical beliefs. It also includes any genetic or health-related data and biometric data for identification; or data concerning a person’s sex life or sexual orientation. While some of these sensitive data might, on occasion, be useful in disambiguation, the costs in complying with the law and the potential ethical jeopardy would almost certainly outweigh the benefits for collections. Therefore, we recommend that sensitive data are not used by collections for disambiguation and should not be stored in collection databases. There are occasions when such data are in the public domain but, if such data need to be stored for research purposes, they should be done under the supervision of the ethical policy of the research institution. Other personal data of living people, such as email addresses and birth dates are covered under GDPR legislation and may be stored in collection management systems, but they cannot be shared without consideration of the regulations.

Outside of Europe, each jurisdiction has different regulations governing the collection and use of personal data. In many cases, laws around the disclosure of personal information echo the intent of GDPR's legitimate interest; however, to our knowledge, these principles remain untested in relation to natural history specimens and it is beyond the scope of this paper to provide guidance on privacy law. It is important that data managers are cognisant of local legislation and the implications of sharing data beyond legal jurisdictions. In addition to legal protections on the use of personal data, specific codes of ethics have been drawn up for museums (Lewis 2004, International Council of Museums 2017, Macdonald 2018, Rabeler et al. 2019). Yet, it is important to be aware that rapid advances in law and technology have changed the ways collections operate and codes of conduct and institutional policies have not necessarily kept pace. Researchers and institutions need to be vigilant of the changing legal, technical and ethical landscape and change their practices where necessary.

Cultural considerations

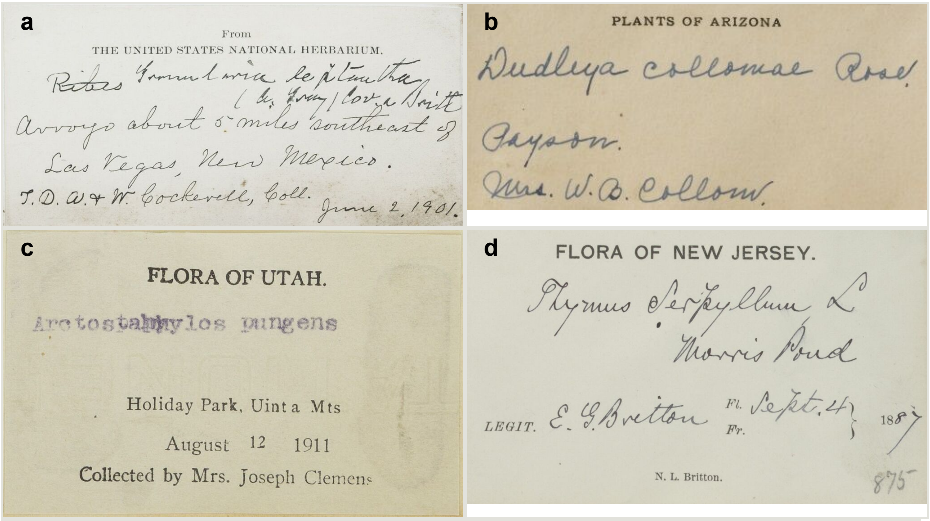
There are many cultural differences that influence how people's names are constructed and used and this makes their disambiguation even more complex. The Anglo-Saxon traditional sequence of title, first name, middle name/s, family name/s, suffix is only one of many ways in which names are constructed (see case studies for examples). Software that is constrained to one cultural norm can lead to the truncation and misspelling of names from other cultures. For example, diacritics and ligatures are partly cultural and, in digital files, are partly based on the age of the database. Before Unicode (i.e. pre-1987), software predominantly used ASCII text that severely limited the available characters; we are still living with the legacy from this period. Variations in the ability of different Collection Management Systems to accommodate diacritics and ligatures can result in multiple versions of an individual's name. This in turn amplifies the need for disambiguation, particularly in databases that have been migrated to newer systems. Simplifications of names due to their encoding can itself lead to mistakes and potentially offend. This is a large subject, but fortunately the W3C organisation has a detailed document on it (Ishida 2011).

Prejudices & biases

A significant benefit of disambiguating people's names and identities is that it can help give recognition to women in the sciences whose contributions to biodiversity and botanical knowledge have historically been pushed to the margins and under-recognised (Shteir 1996). In a Western European and North American context, socio-cultural barriers devaluing women's knowledge of the natural world resulted in either a lack of adequate contribution or no contribution at all (George 2006, Lindon et al. 2015, Meeker and Szabari 2020). Specifically, the cultural norm where women are identified in reference to their husbands (e.g. Mrs. Joseph Clemens) can make it hard to unambiguously identify women collectors (Fig. 1) (also see case studies below). Moreover, data entry protocols can serve to further obscure their contributions to natural history by making the women findable only via their husband's name (Maroske and Vaughan 2014).

The informatics landscape of disambiguation

The mass digitisation of specimens and the open sharing of data have completely changed the potential for disambiguation globally. In addition to this, the availability of open editable identifiers has accelerated the processes many fold. Here, we introduce the core elements of the disambiguation landscape: Wikidata, ORCID and Bionomia as a means to link them.



From
THE UNITED STATES NATIONAL HERBARIUM.
~~Ribes~~ Grossularia leptantha
(A. Gray) Cov. & Britt.
Arroyo about 5 miles southeast of
Las Vegas, New Mexico.
T.D.A. & W. Cockerell, Coll. June 2, 1901.

PLANTS OF ARIZONA
Dudleya collomae Rose
Payson.
Mrs. W. B. Collom.

FLORA OF UTAH.
Arctostaphylos pungens
Holiday Park, Uinta Mts
August 12 1911
Collected by Mrs. Joseph Clemens

FLORA OF NEW JERSEY.
Thymus Serpyllum L.
Morris Pond
LEGIT. E. G. Britton Fl. Fr. Sept. 4, 1887
N. L. Britton.
875

Figure 1.

Examples of labels from herbarium specimens where the female collector has gone unrecognised.

a: The label of a specimen of *Ribes leptanthum* A.Gray collected by Theodore Dru Alison Cockerell and Wilmatte Porter Cockerell from The New York Botanical Garden Herbarium (NY) where only her husband was documented. URI:http://sweetgum.nybg.org/science/vh/specimen_details.php?irn=4063571. Catalogue Number 3771543 (©The New York Botanical Garden, CC-BY-4.0). doi

b: The label of a specimen of *Dudleya collomiae* Rose collected by Rose Eudora Collom from the Royal Botanic Gardens, Kew where the loss of the prefix “Mrs” has led to the specimen being mis-attributed to her husband W. B. Collom. URI:http://specimens.kew.org/herbarium/K000838434. Catalogue Number K000838434 (©Royal Botanic Gardens, Kew, CC-BY-4.0). doi

c: The label of a specimen of *Arctostaphylos patula* Greene collected by Mary Strong Clemens from California Botanic Garden Herbarium, where the loss of the prefix “Mrs” has led to the specimen being mis-attributed to her husband. URI:https://cch2.org/portal/collections/individual/index.php?occid=3895834. Catalogue number RSA0178692 (©California Botanic Garden Herbarium, CC BY-NC-SA). doi

d: The label of a specimen of *Thymus serpyllum* L. collected by Elizabeth Gertrude Britton in the collection of the National Museum of Natural History, Smithsonian Institution Herbarium, but attributed to her husband Nathaniel Lord Britton in the collection of the National Museum of Natural History, Smithsonian Institution. URI: http://n2t.net/ark:/65665/32616a5e9-2143-490a-a0b3-a5d2546376de. Catalogue number US 132144Z (©Smithsonian National Museum of Natural History, CC-0). doi

Wikidata

Wikidata has revolutionised our ability to disambiguate people: it contains only public domain data; it can be edited by anyone; and it is readable by both people and machines. Wikidata is not a primary source for biographical data; indeed, each entity is expected to be referenced to a primary source. Wikidata provides a globally unique and resolvable URL identifer by assigning a "Q identifier" for each person (e.g. the phycologist Josephine Tilden Q20856036). Structured statements describe detailed characteristics of an "Item", such as a person, through community-agreed sets of properties (e.g. the property for "date of death" is P570), values (e.g. 1955-01-12) and supporting evidence – the reference to the primary source (e.g. a URL to an obituary). Statements are created by human volunteers as well as authenticated scripts, known as bots. All data are dedicated to the Public Domain and may be freely accessed through a web interface or through scripted SPARQL queries. Many properties exist for external identifiers (e.g. DOI = P356, ORCID = P496) and, as such, Wikidata is a powerful tool that may be used to store and connect all existing identifiers and to reconcile, broker and resolve entities using their previously disconnected identifier systems.

A Wikidata entry not only provides core biographical details of people, but it also creates a bridge between different sources of authority, such as VIAF and ORCID (Fig. 2). By using Wikidata as a tool for disambiguation and contributing disambiguated data back into it, we can create open public records for people and link them to other resources which bring together all the public information about that person. Traditionally, disambiguation work would have only been documented within a local collection management system available only to local users and few collection management systems have supported capabilities to perform and document such disambiguation work. Now, with Wikidata, important elements of biography can be made available to the whole community. As Fig. 2 illustrates, Wikidata brings together biographical information from the different disciplines of biology. There are some identifiers for specific groups of people, such as botanists, entomologists and zoologists, but it is also clear that there is large overlap between the people in all of these different sources.

Wikidata, like Wikipedia, has the advantage of being community-edited. It is a centralised resource and ensures that data on contributors to collections are openly linked and can be easily enriched and corrected. The notability criteria for the creation of a Wikidata item are significantly lower than that for the creation of a Wikipedia article (see case study below on Winifred Chase); a person is notable for Wikidata if they "*can be described using serious and publicly available references*" (https://www.wikidata.org/wiki/Wikidata:Notability). Wikidata is also a multilingual project allowing editors to contribute in their preferred language.

ORCID

An ORCID (Open Researcher and Contributor IDentifier) is a globally unique, persistent identifier that is free of charge. It is an essential tool to unambiguously identify researchers when submitting manuscripts for publication, grant applications and other scholarly activities (Haak et al. 2012). ORCID makes its metadata publicly available through an application programming interface at the full discretion of each researcher. ORCID is a global, not-for-profit organisation that is supported by membership fees. It is governed by a Board of Directors with representation from the library, publications and research communities. Unlike Wikidata Q Identifiers, which can be created and edited by anyone, ORCID profiles are established and maintained by the individuals they represent. This means ORCIDs have two major limitations:

1. they cannot be used for people who are unable (i.e. the dead or infirm) or unwilling to create one and
2. they often contain too little information to make them useful for disambiguation; it is possible to create a profile with no information at all and for the owner to make their profile private. A detailed and publicly available profile is the best aid for disambiguation.

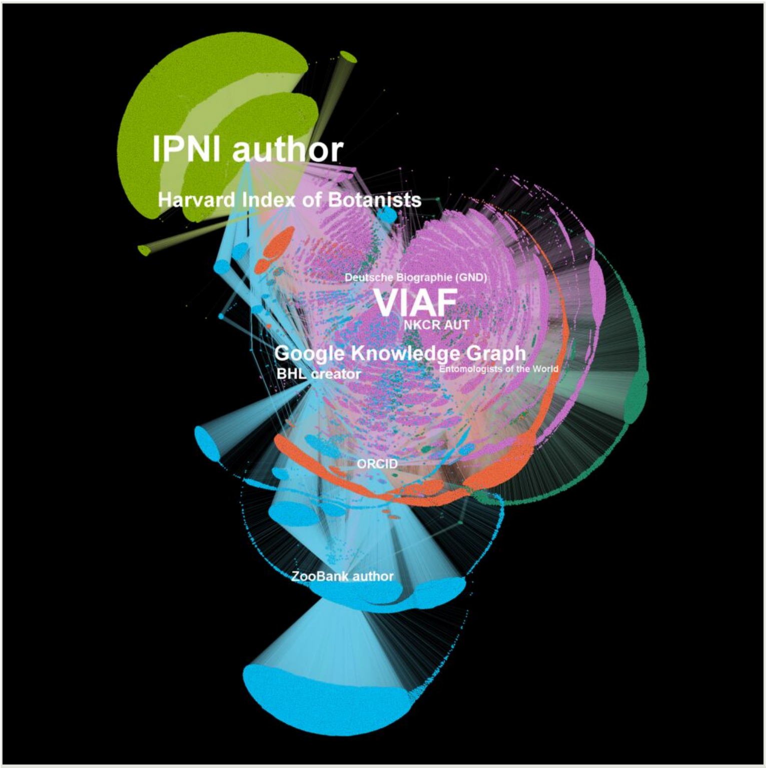

IPNI author
Harvard Index of Botanists
Deutsche Biographie (GND)
VIAF
NKCR AUT
Google Knowledge Graph
BHL creator
Entomologists of the World
ORCID
ZooBank author

Figure 2. doi

A network of the top twenty most used identifiers for biologists on Wikidata. Ten of the more distinct identifiers are labelled with the font size proportional to the number of linked people. The other identifiers are the Dutch National Thesaurus for Author names, the Center of Warsaw University Library catalogue, International Standard Name Identifier, German National Library ID, Library of Congress authority ID, Bibliothèque nationale de France ID, the identifier for authority control in the French collaborative library catalogue, Freebase ID and WorldCat Identities ID, all of which cluster closely together with the VIAF ID due to the large amount of redundancy across those databases. Lastly, the botanist author abbreviation clusters with the IPNI author because there is a one-to-one relationship between these two identifiers. These data were extracted from Wikidata on 18-02-2022. They were visualised using Gephi (Bastian et al. 2009) using the Yifan Hu layout algorithm (Hu 2005) and the five coloured modules are identified using the algorithm of Blondel et al. (2008).

Bionomia

Bionomia is an online, open data curation tool for the disambiguation of collectors and determiners of specimens (Shorthouse 2020). Occurrence records are periodically downloaded from the Global Biodiversity Information Facility (GBIF) then re-indexed and re-organised according to the likelihood that each specimen, as a component of the occurrence, was collected and/or identified by a particular agent. Users of Bionomia then actively verify or reject these candidate, algorithmically-produced assertions. The infrastructure uses a combination of name parsing to split lists of agent name strings into component parts, graph theory algorithm to identify and rank agent name strings and search logic to facilitate human judgement. Participants assert associations between GBIF occurrence records and people as these are represented by their presence in ORCID or Wikidata. As of July 2022, 1,825 participants have either claimed or attributed 22M occurrence records to either an ORCID or a Wikidata Q Identifier, thus persistently linking them to a uniquely identified individual. Collection curators are encouraged to use the data from Bionomia to enrich their collection management system by downloading relational, Frictionless Data packages, incorporating the asserted links, then republishing their data to GBIF with the inclusion of these new links in the Darwin Core fields recordedByID and identifiedByID. GBIF is working with Bionomia to raise awareness and make it easier to bring annotations into collection management systems, similar to other data quality flags.

Bionomia is one example of an online platform where person data participate in a roundtrip (Fig. 3). Other domains have also explored data roundtripping, most notably in libraries, archives and art museums (Larsson et al. 2019, Triebel and Scholz 2022, Agenjo-Bullón and Hernández-Carrascal 2020). There are many semi-structured data elements in collection management systems that could benefit from the application of shared, unique identifiers in open data curation environments with roundtrips back to their source. Geographic place names, scientific names and concepts and bibliographic references are prime candidates to help democratise the exchange of metadata and information through unified services that increase efficiencies, deduplicate effort and engender new collaborations (Nicolson et al. 2018, Güntsch et al. 2021).

Sustainability

The resources and infrastructure of disambiguation need to be sustainable. Wikidata and ORCID have funding strategies and governance models to ensure their longevity. As the museum and herbarium community do not have the funds and capital to replicate these resources within their own domain, there is much value in supporting these common open resources. Bionomia is presently in the start-up, exploratory phase of its development, but is, nonetheless, open by design. It aids disambiguation through stand-alone libraries of code, reusable search algorithms, user-driven exports to digital archives and wholesale downloads that contain annotated links using open standards. Sustainability is less of an issue for Bionomia because it is just a tool, albeit a very useful one, which would eventually become redundant once collections are completely disambiguated.

There is an opportunity now for the community of biological collections to recognise the great value gained from these resources and to engage with them to help steer their development and to contribute data. Ultimately, open informatics resources survive if they are useful to someone and, thereby, give sufficient value to gain investment. We believe this will be the case for these resources and, as more collections join this initiative, the more permanent these resources will be.

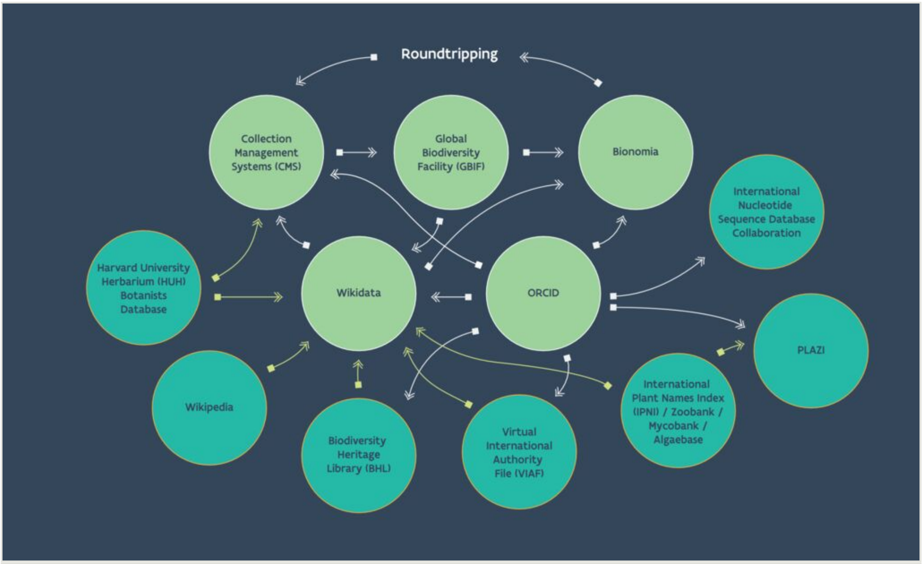

Roundtripping
Collection Management Systems (CMS)
Global Biodiversity Facility (GBIF)
Bionomia
International Nucleotide Sequence Database Collaboration
Harvard University Herbarium (HUH) Botanists Database
Wikidata
ORCID
PLAZI
Wikipedia
Biodiversity Heritage Library (BHL)
Virtual International Authority File (VIAF)
International Plant Names Index (IPNI) / Zoobank / Mycobank / Algaebase

Figure 3. doi

The connections between the core platforms of open disambiguation for natural history collections (light green) and their connections with other important biodiversity informatics platforms (blue). The diagram shows how data flow from collection management systems, through GBIF and is connected in Bionomia to people through Wikidata and ORCID. These new links can then be returned to the system. This data enrichment cycle is referred to as data roundtripping. Other biodiversity informatics platforms facilitate this process by providing biographical data and other information that support disambiguation.

Relevant informatics resources

In additional to the other informatics resources, we mention two others which deserve particular attention, these being the exchange data standards and the tool OpenRefine.

Exchange standards

Darwin Core and ABCD are the two primary standards for exchanging data on biological specimens (Wieczorek et al. 2012, Holetschek et al. 2012). Recently, Darwin Core added the terms recordedByID and identifiedByID to allow the addition of person identifiers to the terms recordedBy and identifiedBy. Additionally, the latest ABCD release version 3.0 is able to accommodate identifiers for people (Access to Biological Collection Data task group 2019). There are other roles for people involved in the collection, identification and study of specimens. Currently, the People in Biodiversity Data Task Group of the Biodiversity Information Standards organisation are working on an extension to Darwin Core Archive exchange format that will expand the range of options for documenting information about people and their roles. This will include name strings and identifiers for multiple people in ranked order. Nevertheless, even this extension to the standard is not intended to handle multiple identifiers for the same person and the complexities of verbatim name strings, versus canonical names. Such details will still be important in biographical databases and collection management systems.

OpenRefine

OpenRefine is a browser-based tool for cleaning, transforming and enriching tabular data. OpenRefine allows users to import spreadsheet style data, which can then be reformatted using a simple expression language known as GREL, built to resemble JavaScript. Alternatively, more technical users can also write expressions in Jython (a Java implementation of Python) or Clojure. Importing libraries from the last two languages greatly extends the possibilities for transforming data within OpenRefine, for example, the processing of XML/HTML elements via jsoup. This also allows linkage to web services, whereby matches to potential names can be made and users can select the disambiguated name (van Hooland et al. 2013). OpenRefine is particularly well suited to bulk disambiguation of records, because it can apply decisions across multiple rows and implements several fuzzy matching algorithms for record clustering, including phonetic matching. This is especially useful for terms that sound the same, but have alternative spellings, which is sometimes the case for person names. It is well suited to disambiguation after specimen labels are transcribed, but before they are imported into a collection management system. There are several manuals available on how to use OpenRefine, but Hill (2016) and Sterner (2019) were specifically written for people working on collections.

Best practices

The disambiguation process can be visualised as a cycle of disambiguation that enriches and links data through identifiers with the aim of completing a roundtrip of data improvement (Fig. 4). The entry point into this cycle may vary depending on the trigger for the disambiguation event and some disambiguation work may also be more focused and, therefore, restricted to one section of the cycle.

The process of disambiguation is inherent to most aspects of working with people names including data capture, management and analysis. For example, the moment a person's name is entered into a collection management system, a decision is made about how that name is recorded and if it should be associated with other names and records in the system or with external resources. Whether entering data for newly-collected specimens or bulk enhancement of historical specimens already held in collections, ensuring that the people's names are unambiguous and, where possible, associated with an identifier, is an important aspect of collection data management.

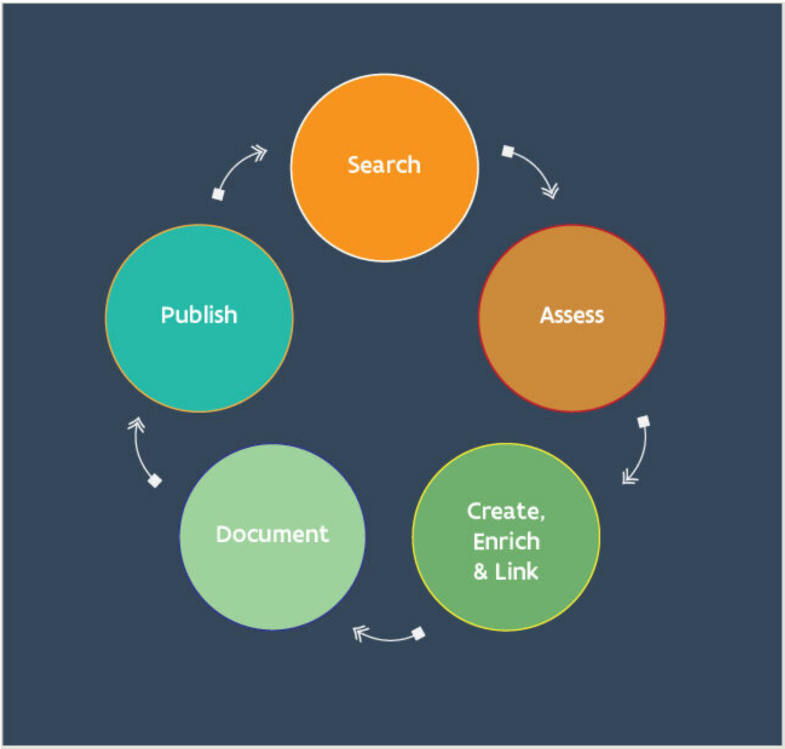


Figure 4. doi

Disambiguation is a cycle. Enrichment of the data feeds off itself leading to further disambiguation. As more names are disambiguated and more biographical data are accumulated, it becomes easier to disambiguate more names.

The process of disambiguation for natural history collection data may be triggered by a wide range of activities undertaken by many individuals including curators, data managers, researchers and citizen scientists. The range of triggers and the kind of activity being undertaken may result in people entering the disambiguation process at various points, carrying out work within all or part of the disambiguation process, with each part of the process being more or less iterative.

In this section, we provide an overview of the steps that may be involved in disambiguating people's names and the individuals to whom they refer, namely: preparation, prioritisation, searching, assessing, creating, enhancing, linking, documenting and publishing. Recommendations for the disambiguation process have been identified and are included within the relevant stages of disambiguation below. In addition, recommendations have been identified for data capture and management; implementing these recommendations will reduce the need for disambiguation in the future.

Before you start

The need for disambiguation could be triggered by any number of activities, events or research needs. For example,

- an institution may decide to systematically disambiguate their collection, perhaps to improve the data quality and findability of their specimens;
- disambiguation might be the focus of a specific project, such as evaluating the contribution of women to taxonomy or studying the history of scientific exploration of a region; or
- disambiguation might be caused by a specific event, such as the death of a collector, the release of data from another collection or a novel data analysis.

Preparation

Depending on the activity and the trigger for disambiguation, it may be helpful to consider some preparatory steps to aid the process. Creating batches or clusters of records that require disambiguation will make the process easier, faster and more accurate. Clustering might be on date, collecting location, the taxonomy of the specimens, co-collectors or anything that will enrich the batches for one or a few people. These aggregated records will often provide additional information and context, as well as providing ranges and variation within the data. Additional disambiguation techniques may also use batches of digital specimen images that are processed through image analysis software (e.g. trained machine-learning models) to extract data specifically for disambiguating the names within the specimen records (Walton et al. 2020).

Recommendations

- Batch or cluster records/images to increase the information and context of each record and to reduce the likelihood of repeating the disambiguation process for related records.

Prioritisation

Determining the most appropriate order to disambiguate will save time, maximise utility of the data and reduce the overall effort. Disambiguation work will be most efficient if it is focused on a particular taxon or geographic area. Working on subsets of clustered data provides clear boundaries around the people involved and their co-collectors, identifiers and publications. If a more general disambiguation is envisaged, then prioritising the most frequently occurring names means a large proportion of specimens will be resolved quickly. On average, if you can disambiguate 3% of the most prolific collectors or identifiers, this will connect those people to 80% of specimens in a collection (Groom et al. 2020).

In some cases, a person's name is relatively distinctive, at least within a collection. Names such as these may be possible to link to an identifier directly through automatic string comparison. Additionally, as disambiguation supports further disambiguation, roundtripping of identifiers into collection management systems and other databases helps to simplify and accelerate further disambiguation.

There is also a good case to prioritise the disambiguation of names for which undocumented knowledge is available, such as for people who are alive, recently died or for whom an oral history still survives.

It should be remembered, however, that due to historic data practices, many of the participants of biological collections have been under-reported. Collection should not reinforce this discrimination by always focusing on the most conspicuous people, even if this represents the easiest path.

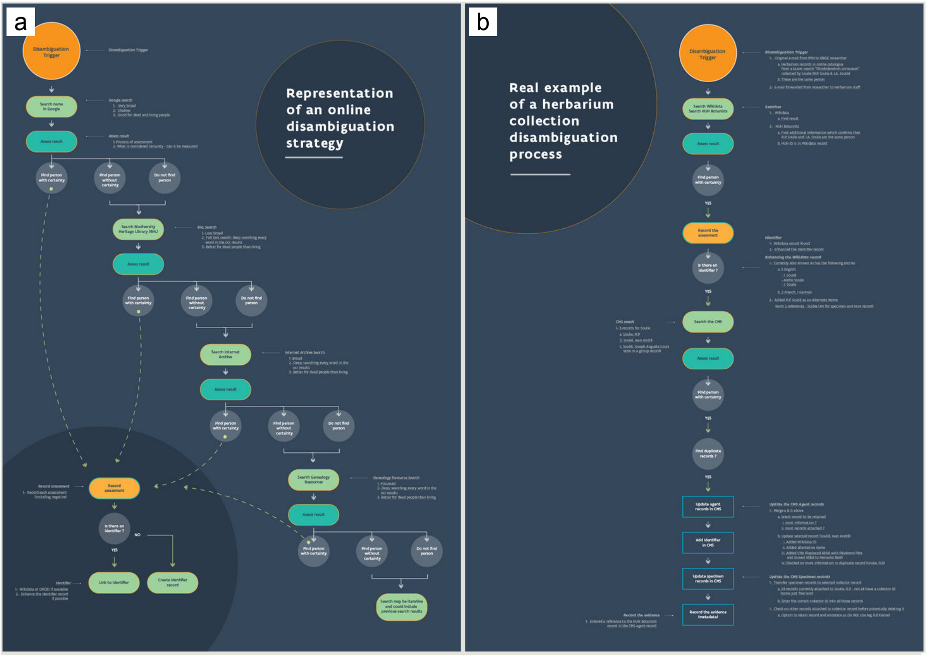

a
Disambiguation Trigger
Disambiguation Trigger
Search name in Google
Google search
1- Very broad
2- Shallow
3- Good for dead and living people
Assess result
Assess result
1- Process of assessment
2- What is considered certainty - can it be measured
Find person with certainty
Find person without certainty
Do not find person
Search Biodiversity Heritage Library (BHL)
BHL Search
1- Less broad
2- Full text search: deep searching every word in the ocr results
3- Better for dead people than living
Assess result
Find person with certainty
Find person without certainty
Do not find person
Search Internet Archive
Internet Archive Search
1- Broad
2- Deep, searching every word in the ocr results
3- Better for dead people than living
Assess result
Find person with certainty
Find person without certainty
Do not find person
Search Genealogy Resources
Genealogy Resource Search
1- Focussed
2- Deep, searching every word in the ocr results
3- Better for dead people than living
Assess result
Find person with certainty
Find person without certainty
Do not find person
Search may be iterative and could include previous search results
Representation of an online disambiguation strategy
Record assessment
Record assessment
1- Record each assessment (including negative)
Is there an identifier ?
NO
YES
Link to identifier
Create identifier record
Identifier
1- Wikidata or ORCID if available
2- Enhance the identifier record if possible
b
Real example of a herbarium collection disambiguation process
Disambiguation Trigger
Disambiguation Trigger
1- Original e-mail from IPNI to RBGE researcher
a. Herbarium records in online catalogue from a taxon search "Rhododendron coriaceum" collected by Soulie (R.P. Soulie & J.A. Soulie)
b. These are the same person
2- E-mail forwarded from researcher to Herbarium staff
Search Wikidata Search HUH Botanists
Searches
1- Wikidata
a. Find result
2- HUH Botanists
a. Find additional information which confirms that R.P. Soulie and J.A. Soulie are the same person
b. HUH ID is in Wikidata record
Assess result
Find person with certainty
YES
Record the assessment
Identifier
1- Wikidata record found
2- Enhanced the identifier record
Enhancing the Wikidata record
1- Currently Also Known As has the following entries
a. 3 English
- J. Soulié
- Andre Soulie
- J. Soulie
b. 2 French, 1 German
2- Added R.P. Soulié as an Alternate Name (with 2 references - Stable URI for specimen and HUH record)
Is there an identifier ?
YES
CMS result
1- 3 records for Soulie
a. Soulie, R.P.
b. Soulié, Jean André
c. Soulié, Joseph Auguste Louis (also in a group record)
Search the CMS
Assess result
Find person with certainty
YES
Find duplicate records ?
YES
Update agent records in CMS
Update the CMS Agent records
1- Merge a & b above
a. Select record to be retained
i. most information ?
ii. most records attached ?
b. Update selected record (Soulié, Jean André)
i. Added Wikidata ID
ii. Added alternative name
iii. Added title (Replaced Abbé with Révérend Père and moved Abbé to Remarks field)
iv. Checked no more information in duplicate record (Soulie, R.P.)
Add identifier in CMS
Update specimen records in CMS
Update the CMS Specimen records
1- Transfer specimen records to selected collector record
a. 28 records currently attached to Soulie, R.P. - not all have a collector ID (some just free text)
b. Enter the correct collector ID into all these records
1- Check no other records attached to collector record before potentially deleting it
a. Option to retain record and annotate as Do Not Use (eg R.P. Klaine)
Record the evidence (metadata)
Record the evidence
1- Entered a reference to the HUH Botanists record in the CMS agent record

Figure 5.

Disambiguation strategies. Abbreviations: Harvard University Herbaria Index of Botanists (HUH); International Plant Names Index (IPNI); Royal Botanic Garden Edinburgh (RBGE). Also available in PDF format (Suppl. materials 1, 2).

a: A diagrammatic representation of one disambiguation strategy. Strategies vary considerably depending on the name being disambiguated, the dates involved, the taxonomy of the specimen, the collecting locality and the collection it is held in. doi

b: A real example of how a name string is disambiguated and the steps taken in documenting it. doi

Recommendations

- Prioritise disambiguation activities that have high value for research and that balance representation of collectors.

- Integrate disambiguation processes, activities or outcomes within new or existing transcription activities to maximise efficiencies.

Search

An element of searching is part of most disambiguation processes. Indeed, a search could be the trigger for disambiguation when a name cannot be resolved by the resulting data discovered during the search.

Searching in collection management systems

Undertaking a search in an institutional collection management system is often the first step in the disambiguation process (Fig. 5, also available in a PDF format Suppl. materials 1, 2). The data structure in a collection management system will often have an impact on how an initial search is undertaken and it may be necessary to carry out a search in several tables and fields to ensure that all relevant records and information are found. However, where an individual is recorded in a system under more than one role, such as collector, identifier or staff member, it might be easiest to disambiguate the application of the name by one role at a time. For example, if the name is linked to both the collecting and determining of specimens, query for occurrences of the name as collector and disambiguate those first.

To differentiate between two or more people with the same name, it is useful to sort the occurrence records by different values, such as collecting date, collecting numbers and locality (and, in some cases, co-collectors and taxa). To ensure the results of the search include this information, it may be necessary to define the structure and content of the results within the search itself. Therefore, when searching a collection management system:

- Consider which data are expected in the results and how they will be exported.
- Look for patterns in collecting dates, collecting numbers and/or localities that identify the activities of an individual. Depending on the number of records, it may be helpful to chart collecting dates or map collecting localities to reveal patterns in collecting activity.
- Group records that appear likely to be collected by the same individual by annotating each row or splitting the data into separate datasets. The ability to easily view the range of specimens already collected by an individual to determine whether they were all collected by the same individual or to match a new specimen to the correct collector record in the collection management system will be important for the efficiency and accuracy of any disambiguation process. It will often be necessary to find more information than is available in a system to disambiguate names.
- Try to establish the identities of the different individuals by searching online or in other resources.
- Be mindful that apparent anomalies in collecting localities can sometimes be explained by travel for personal or professional reasons. Knowledge of where a

collector's extended family resided can provide a justification for travel to a given locality, especially for non-professional collectors in the 19th century or earlier.

Recommendation

- Develop a good strategy for searching, with an understanding of the methodology, limitations and context of each collection management system used.

Searching online

Once a target for disambiguation has been identified and any local information, such as related specimens, has been collated, the easiest option is to use a search engine, such as Google and search for the person's name. For notable people, this may be all that is required to establish who they are and link them to an existing identifier. However, such searches often need to be made more specific by the addition of biographical material, such as the name of an institution where they worked, their birth or death year or the name of someone they worked with. Such searches will often lead to additional information that can be used to refine searches and/or query databases, such as the Biodiversity Heritage Library, the Internet Archive, FamilySearch, Ancestry.com, bibliographic databases, Wikipedia and Wikidata (see the Sources of Information section below).

Each of the resources mentioned above differs in its scope and limitations. A person may have an extensive record in one dataset, but be missing entirely in another. Or, for example, Google Scholar might retrieve very different results to Biodiversity Heritage Library full text search. Search results from different datasets will often complement each other and, as biographical and specimen data are continually growing in availability on the internet, searches can always be revisited with renewed chances of success.

Recommendations

- Develop an online search strategy with documented methodology that is sensitive to the limitations and context of each online resource used.
- Become familiar with candidate search resources and use them in concert to refine your evidence.

Assess

The assessment of a manual disambiguation process is based on the experience of the person carrying out the disambiguation. Experience only comes with practice and learning and even very knowledgeable disambiguators can be misled by the data. It is important to judge whether, on the balance of probability, the existing data on both the specimen and the person are sufficient to make the link between these two entities.

The criteria used might include dates, places, taxa, handwriting, label format and co-collectors, but the weight placed on each data type is up to the disambiguator and their knowledge of the specimen and the person they want to link to it. Disambiguators need to be self-critical and be willing to re-evaluate and return to decisions in the light of new information. It is important to document the data as they are uncovered, but it may not be possible to document the complete trail of breadcrumbs that often precedes the discovery of a person and their biography (Fig. 5, Suppl. materials 1, 2).

Recommendation

- Be prepared to accept failure. Disambiguation is an iterative, cumulative and re-affirmative process, but it is not absolute.

Create, enhance and link

When disambiguating the names of people attached to specimens, it may be necessary to create new records in a collection management system. It may also be possible to enhance existing records, based on new information gained during the disambiguation process. When creating a record in a collection management system, there may be specific protocols and constraints on how the data should be entered. The use of verbatim fields can provide useful information on how a person's name has been written on specimen labels that can aid future disambiguation. Verbatim fields can also be used to record names that cannot be unambiguously assigned to a single individual.

Where possible, enhance the record for an individual with information that will aid the correct use of the record in future. If you have specimens that have been collected by the individual, then it is usually possible to provide dates which bound the period of activity of that person (i.e. floruit dates, sometimes abbreviated to fl.). Likewise, the geographical region and taxonomic interests of a person can be determined and added to the biographical record. The disambiguation process will also often require new identifier records to be created in the authority resources. When creating or enhancing a record in Wikidata, aim to include at least one reference. If you are creating a Wikidata record for someone for whom you have very little information, then the reference could cite a specimen collected by them.

Linking the local record in a collection management system with the identifier record in the authority resource is key to embedding the disambiguation into the data. The process of creating the link may vary depending on the collection management system and the resource. If there is a central table of people records within the system, this may be the most appropriate place to hold the Wikidata or ORCID identifier (the link to the global resource). The structure and functionality provided by the link may also impact the level of information stored locally in the system. Some locally maintained data may be the preferred option if the local system is quite isolated or there are sensitive data involved.

Recommendations

- Capture the person's name verbatim from the label in the specimen record and treat disambiguation as a secondary enrichment process.
- Link to and contribute to large communal biographical resources, such as Wikidata, rather than large local stores of biographical information.

- If an identifier does not exist for a person, create one, ideally a Wikidata Q Identifier with a minimum of one reference.
- If an identifier already exists, check whether any information gained from the disambiguation process can enhance the authority identifier record.

Document

If a disambiguation decision has not been thoroughly documented, explained and referenced, there is a risk it may be undone in the future. It should be considered how the decision should be recorded. That is not to say that all the information needs to be documented in the same place. Enriching a collection management system is suitable for specimen related annotations; however, biographical information may be better suited to be documented in Wikidata or even Wikipedia, if the person is notable enough.

The person who made the assessment and the date should be recorded; in most systems, this is automatic. Notes, flags and tags all might be useful to document levels and sources of uncertainty. If disambiguation attempts fail, it is equally important to record the reason and information discovered in the process. To indicate the reasons for a failed disambiguation, the following flags might be used:

- No biographical details are available for collector's/determiner's name.
- The specimen has insufficient information to attribute to a collector/determiner.
- There are candidate people, but it is suspected that there is an error on the specimen label or in the biography of potential collectors/determiners that prevent disambiguation.

Sometimes, names written on specimens cannot be disambiguated. For example, it can be difficult to separate a husband and wife who often travelled together and have the same initials. In some collection management systems, it might be possible to create a collecting team, but it is not appropriate to record such a team in Wikidata. It is not uncommon to fail at disambiguation, but to have enough information to limit the choice to a small number of people. The results should then only be recorded at the specimen level, although information relating to the potential confusion of individuals could be documented in the person records of a collection management system.

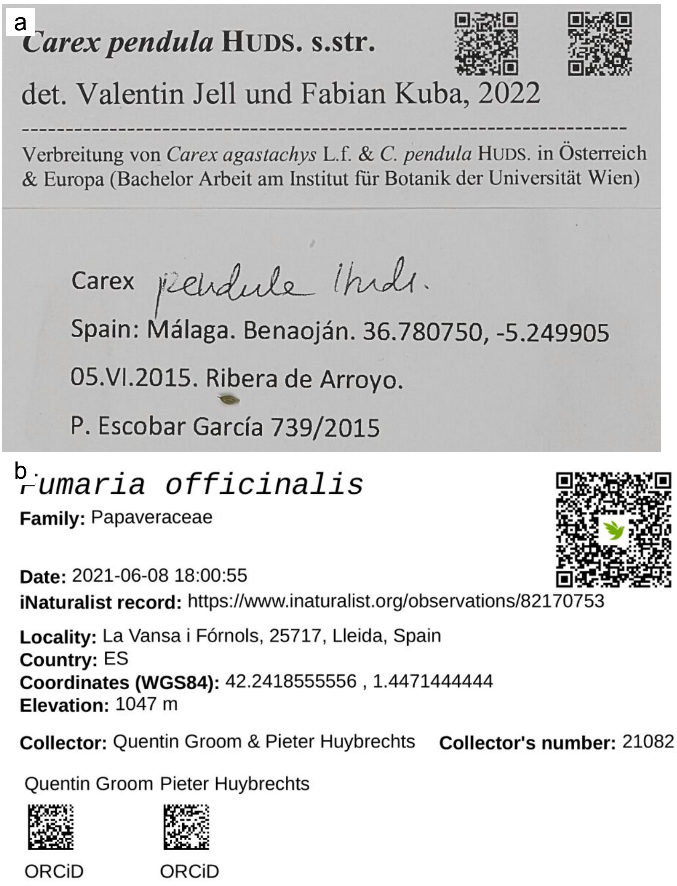
Recommendation

- Record the results of the disambiguation whether or not success was achieved.
- Enrich public domain resources, such as Wikidata, because this benefits global disambiguation activities.

Publish

Institutional collection management systems contain a wealth of carefully-curated data researched by people with extensive specialist knowledge. This may include years of work ensuring that the person names are unambiguous. If this work is not published, this information will not be made available to the wider community. Publication processes in the past have often made it difficult to include this information, particularly when submitting data to aggregators.

However, the addition of the Darwin Core fields “recordedByID” and “identifiedByID” mean that the identifier for the collector and the determiner can (and should) be included with published data. These terms are also available for use by institutions, enabling the inclusion of person identifiers in institutional portals. These identifiers can, thus, be included in data downloads and on printed specimen labels (Fig. 6).



Carex pendula HUDS. s.str.

det. Valentin Jell und Fabian Kuba, 2022

Verbreitung von *Carex agastachys* L.f. & *C. pendula* HUDS. in Österreich & Europa (Bachelor Arbeit am Institut für Botanik der Universität Wien)

Carex pendula Huds.

Spain: Málaga. Benaoján. 36.780750, -5.249905

05.VI.2015. Ribera de Arroyo.

P. Escobar García 739/2015

Fumaria officinalis

Family: Papaveraceae

Date: 2021-06-08 18:00:55
iNaturalist record: https://www.inaturalist.org/observations/82170753

Locality: La Vansa i Fórnols, 25717, Lleida, Spain
Country: ES
Coordinates (WGS84): 42.2418555556 , 1.4471444444
Elevation: 1047 m

Collector: Quentin Groom & Pieter Huybrechts **Collector's number:** 21082

Quentin Groom Pieter Huybrechts

ORCiD ORCiD

Figure 6.

Two examples of labels where either the determiners (A) or collectors (B) have been unambiguously identified by their ORCID, encoded into QR codes (A) or data matrices (B) on the label.

a: From specimen https://w.jacq.org/W0132025 (©Natural History Museum, Vienna, CC-BY 4.0). doi

b: From Meise Botanic Garden (CC-0). doi

Recommendations

- When publishing a person's name, include the full name and any titles and suffixes.
- Publish the identifiers attached to the records when publishing data on portals or to aggregators.
- Where possible, include the person identifier for the collector on the labels for new specimens. This might be done as part of the accessioning process.

Sources of information

People and their lives can be described through various types of data, which can be obtained and cross-referenced from multiple sources. Below is a non-exhaustive list of the different characteristics that may be leveraged for disambiguation. These are sometimes only understandable in the context of the era and culture of the people to whom they refer. The person doing the disambiguation must make a judgement as to which sources are likely to provide fruitful results.

- Depictions: portraits, photographs
- Physical traits: ethnicity, sex, other distinguishing features
- Identifiers: names, aliases, signature, persistent URIs
- Linguistics: language, dialect, vocabulary
- Interests: taxonomic groups, localities
- Awards: prizes, grants, titles
- Positions/roles: jobs, qualifications, memberships
- Birth-death: dates of, active lifespan, attendances
- Proxies: parents, children, siblings, spouses, friends, colleagues, students, supervisors, co-authors, co-collectors
- Material evidence: publications, presentations, specimen labels, letters, field books, certificates, collection registers

Below, we outline the most relevant resources of biographical information for biological collections. There are many more. With experience, disambiguators will discover suitable sources for the category/ies of people they frequently work on. The lengths one is prepared to go to disambiguate someone depends on the nature of the project, the importance of the person and the likelihood of success.

Genealogical websites

The genealogical community and their research are an extremely helpful resource for disambiguating people. The decades of work by this community have culminated in multiple websites that contain a wealth of interconnected data on people, including family relations, birth, death and life event dates, as well as links to primary source documentation that support those data. Examples of these websites include Ancestry.com, Billiongraves.com, Familysearch.org, Findagrave.com, Findmypast.com, Geni.com, Myheritage.com and Wikitree.com to name but a few. The data and linking contained in these websites can greatly assist with the disambiguation of people associated with natural history specimens.

Miscellaneous publications, catalogues and field notes

Co-collectors are often co-authors of scientific publications. A single collector's surname can be difficult to disambiguate, but a pair of surnames is often unique. Resources such as Google Scholar facilitate searches across scholarly publications for pairs or teams of authors. Once potential matches have been found, the publications themselves often reveal more information about the collectors, including their full names and institutional affiliations. The field notes and diaries of some collectors may also be publicly accessible online. *Indices Collectorum* are another source of data. These are a published catalogue of a collector or several collectors' activity. They may be associated with *exsiccata* that may have once been offered for sale (Triebel and Scholz 2022). The Biodiversity Heritage Library (see below) is an excellent source of these. Such resources often contain additional information about expeditions and the collectors themselves. These notes also link specimens together, potentially making them suitable for bulk disambiguation of numerous specimens.

Biodiversity Heritage Library (BHL)

BHL is the world's largest online repository of biodiversity literature and archival materials. It is a global consortium of over 500 libraries and publishers, who have together made over 60 million pages freely accessible online. Users can search the contents of the Library in two ways: by searching the catalogue for publications (and filtering by author, date or subject) or via full-text search, which searches the OCR text across all 60 million pages. This opens up information on little known people and provides valuable biographic references useful for disambiguation. The knowledge gained can add to the totality of evidence and improve the confidence in the disambiguation. BHL also shares its contents with the Internet Archive, which provides yet more content that might have relevance to disambiguation.

Taxonomic Literature II (TL-2)

TL-2 is a guide to the literature of systematic botany published between 1753 and 1940 and was originally a print series in fifteen volumes published by the International Association for Plant Taxonomy (IAPT). A digital version of TL-2 has been made available online by the Smithsonian Institution Libraries. It contains detailed biographies of taxonomists, including their publications, their employment history and the institutions where their specimens have been deposited. It is an invaluable source of information about people publishing in a specific field within a specific time period.

Specimens

Many institutional collection databases have an online portal and some have associated biographical information. The Global Biodiversity Information Facility (GBIF) aggregates biodiversity data from institutions across the globe, thus making it accessible and discoverable within a single portal. The digital occurrence data records, inferred from specimens and other material samples available on GBIF, are limited by the extent of documentation by the data provider for the collection. However, images of specimens from these host collections can be a useful resource in their own right. The specimen labels captured in collection images can contain critical information that has not been transcribed or included in collection data, such as "Mrs" or Jr" (see Prejudices & Biases section above). Specimen labels can also be used to verify spellings of transcribed names and other data.

Wikipedia

Wikipedia is an openly-licensed online encyclopaedia to which anyone may contribute. Wikipedia content is also indexed and ranked highly by search engines. This ensures that, if an article on a collector exists, that article will be one of the first returned search results. Wikipedia has more than three hundred language versions, with the English language version being the largest. These different language versions have independent content, so information missing in one language version may be present in another. Helpfully, Wikidata provides a bridge between the different language versions. Wikipedia collates knowledge about people in an accessible, editable, online resource and is, therefore, of great assistance in disambiguation efforts. As Wikipedia is a centralised resource, any contributions made to Wikipedia are more visible and are likely to have more impact than contributions made to more specialised platforms. Nevertheless, not all people can be included within Wikipedia due to the encyclopaedia's notability criteria. In Wikipedia, people are presumed notable if they have received significant coverage in multiple published secondary sources that are reliable, intellectually independent of each other and independent of the subject. Many contributors to natural history collections fail to meet these criteria, even if they meet the criteria for Wikidata (see Wikidata section above).

Virtual International Authority File (VIAF)

VIAF is a large authority file created from a consortium of international libraries who contribute their own local authority files. VIAF consolidates authorities from its sources and, where possible, aggregates them under a single VIAF ID related to a single person. However, where biographies have not been linked, there may be multiple VIAF IDs for a single person. VIAF is run by OCLC (the Online Computer Library Center), a non-profit membership organisation. As Fig. 2 demonstrates, VIAF plays an important role in the landscape of identifiers, but it is far from a universal identifier for biologists.

International Standard Name Identifier (ISNI)

ISNI is a person identifier for anyone involved in the production of creative works. This includes authors, artists, musicians and their producers and publishers. ISNI is also an ISO Standard Identifier (International Organization for Standardization 2017). This means it has a larger scope than VIAF, but in the context of biological collections, the people included are mainly the authors of publications on science and natural history.

Other identifiers

Other identifiers useful to collection disambiguation work include those in databases, such as Harvard Index of Botanists, International Plant Names Index (IPNI), ZooBank, Biodiversity Heritage Library and Wikispecies. These have an advantage over those in Wikidata, ORCID, VIAF or ISNI in that the person records present in them are more likely to be linked to specimens in a collection. Most of these other identifiers are also available as Wikidata properties and, therefore, Wikidata can also be used to reconcile identifiers between these databases (Fig. 2).

Case studies

Here, we describe some specific disambiguation projects conducted by the authors. These case studies illustrate some of the problems, processes, sources of information and benefits of disambiguation.

Collaborative disambiguation of North American bats

In a project to enhance currently published data for the rhinolophid and hipposiderid bats, some authors of this paper held a workshop with bat researchers and collection managers on 1 December 2020 (Mast et al. 2021). Our goal was to verify the identity of historic and present-day people who collected or identified specimens of these taxa. During this disambiguation marathon, workshop participants uncovered problems, solved them in a virtual group setting and came to appreciate the value of this exercise. Below, we outline some of their discoveries.

Using Bionomia to explore specimens of bats labelled with the seemingly unique string " *Geoffroy Saint-Hilaire*", workshop participants were able to discover patterns and outliers in collection and determination dates which indicated that they were not all collected or determined by one individual. By examining biographical data in Wikidata for Geoffroy Saint-Hilaire, participants discovered that there were, in fact, three people who shared at least parts of the same name: the father Étienne Geoffroy Saint-Hilaire (1772–1844), the son, Isidore Geoffroy Saint-Hilaire (1805–1861) and the grandson Albert Geoffroy Saint-Hilaire (1835–1919) who all collected bats. By cross-referencing birth and death dates, participants were able to decipher which bat specimens were most likely collected by which of these three family members. These deductions later served to help inform a second team of participants who were charged with verifying georeferenced collection localities.

Linking people to specimen records in a publicly accessible, web-enabled environment like Bionomia (and Wikidata) resulted in surreptitious discoveries. Harry Hoogstraal (1917–1986) was an American zoologist and prolific collector of specimens, particularly in the tropics. Prior to the workshop, many of his rhinolophid and hipposiderid bat records had already been linked to him via Bionomia through his Wikidata Q Identifier, Q5669784. As a result of search engines indexing this content, the string "Ibrahim Helmy" was made discoverable on the Internet, plainly seen as a co-collector of Harry Hoogstraal. This was a necessary clue that led to the discovery that Ibrahim Helmy co-authored, "The Contemporary Land Mammals of of Egypt (Including Sinai)", with the late Dale James Osborn, a Research Associate with the American Museum of Natural History (Osborn and Helmy 1980). Ibrahim led field expeditions in support of the US Navy's Naval Medical Research Unit Three (NAMRU-3) Medical Zoology programme then stationed in Cairo, Egypt.

Workshop participants were able to assign ORCID identifiers and/or Wikidata Q Identifier to over 500 people involved in collecting bats. The breadth of these activities revealed unlikely collection dates for some of the earliest-known bat specimens (Fig. 7), which are now under investigation by collection managers. Cross-referencing readily-, openly- and computationally-available birth and death dates of collectors against dates of collection demonstrates the value of involving the research community in this disambiguation activity.

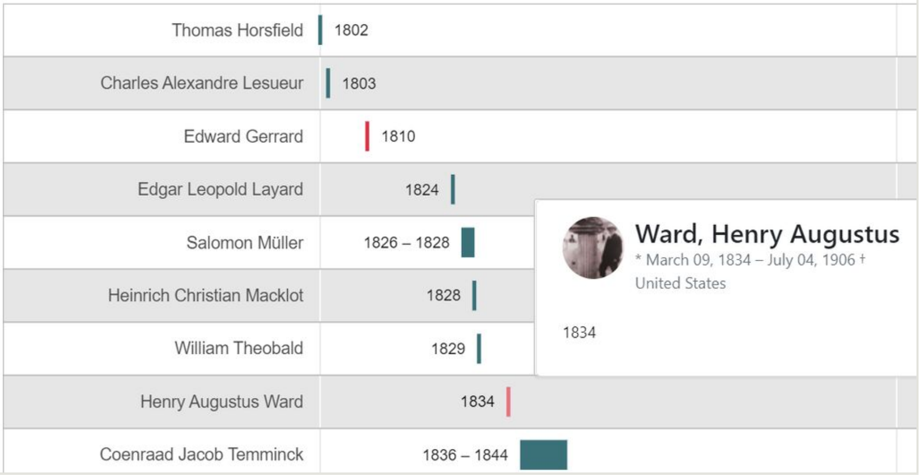


Figure 7. doi

Timeline for the earliest hipposiderid bats (Old World leaf-nosed bats) linked to collectors via Bionomia. The narrow, red bars with the years 1810 and 1834 indicate a problem with these attributed records requiring further investigation. Edward Gerrard was born in 1832 and Henry Augustus Ward was born 9 March 1834.

The Schimpers

The Schimper family of Baden, now part of Germany, produced four important scientists of the 19th century: the brothers Karl Friedrich Schimper (1803–1867) (Fig. 8a) and Georg Wilhelm Heinrich Schimper (1804–1878; GWHS), their cousin Wilhelm Philipp Schimper (1808–1880; WPS) (Fig. 8b) and his son Andreas Franz Wilhelm Schimper (1856–1901; AFWS) (Fig. 8c). Three of them were of the same generation and all made a considerable contribution to our botanical knowledge. The complexity of disambiguating these names when digitising label data from specimens collected by the Schimpers has been outlined in Bräuchler et al. (2021) and is summarised here.

Figure 8.

Pictures of three of the Schimper family, all of whom were productive collectors of herbarium specimens. They are included here to give an example of easily confused names and to remind us that there are people with complex lives behind every name and biography.

a: Karl Friedrich Schimper, by Conrad Geyer (1816-1893), public domain, via Wikimedia Commons doi

b: Georg Wilhelm Heinrich Schimper, public domain from the Bibliothèque nationale et universitaire de Strasbourg doi

c: Andreas Franz Wilhelm Schimper, public domain, via Wikimedia Commons doi

All four Schimpers collected plant specimens that are now in collections around the globe; indeed, many taxa are named after them. Consequently, their names are frequently mentioned in literature and on herbarium specimens. In many herbaria, only the family name is recorded on the labels, so it is not clear which individual collected the specimen. The presence of an initial on the specimen label does little to identify the individual, given that, despite their different forenames, GWHS, WPS and AFWS all went by the name Wilhelm. The resulting confusion has led to two entries for “W. Schimper” in Harvard University Herbaria Index of Botanists (ID 0094171, 0094172). GWHS and WPS are most likely to be conflated, due to their overlapping periods of activity and the large volume of herbarium specimens either collected, identified or distributed by them: both distributed exsiccatae of their collections, either as gifts, exchange or for sale.

Given the wealth of literature and online information on the Schimpers and the sometimes inconsistent information contained therein, it can be time-consuming to disentangle the information needed to disambiguate these collectors.

The collecting locality provides the best starting point for disambiguation, with GWHS alone having collected in Algeria, Greece, Egypt, Saudi Arabia and Ethiopia. Although the country of collection can be used to determine the collector for a large proportion of Schimper specimens, it does not help with specimens from France and Germany, where all four Schimpers collected. For these, a thorough knowledge of the biographies and collecting activities is necessary for disambiguation.

Knowledge of each collector's taxonomic speciality is also useful but, again, there is some overlap: WPS identified and described the mosses collected by GWHS and even distributed part of his Ethiopian specimens via his exchange society, so WPS' name may also be associated with GWHS' specimens.

Again it is clear that high quality biographical data is necessary for disambiguation, but also that data from the specimens themselves contribute biographical information, so that disambiguation benefits from the process of disambiguation itself.

Ethel Winifred Bennett Chase

Ethel Winifred Bennett Chase was an American botanist, a professor of botany and the Dean of Women at Wayne State University in the United States (Fig. 9). Much of her scientifically significant botanical collecting was undertaken on an expedition to the South Pacific, accompanying the algologist Josephine E. Tilden. However, correctly attributing specimens to Chase can be challenging as she preferred to use her middle name, Winifred. The numerous resulting name strings used by herbaria databases have also led to confusion.

In order to assist with disambiguation, a Wikidata item was created for Chase. This item was linked to the item for Josephine E. Tilden through a statement that the two botanists were co-collectors. This ensures that the Wikidata notability criteria was satisfied for both people. Having a Wikidata item allows Chase's multiple aliases to be listed. It also enables the collation of biographical data, institutional identifiers, databases, websites and scholarly articles as supporting references for statements added to that item.

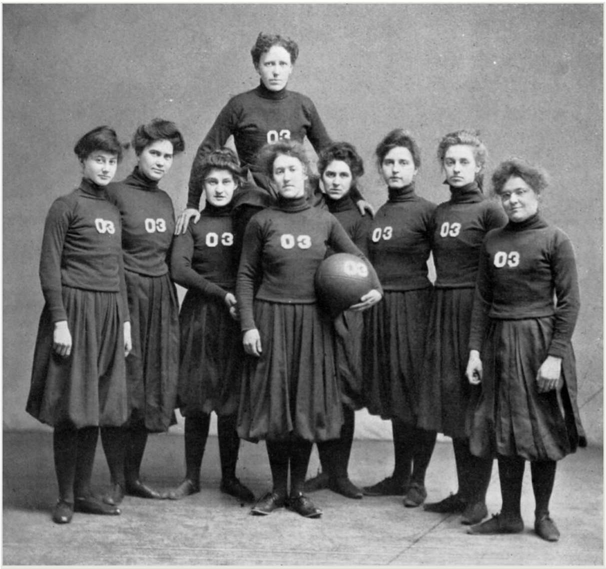


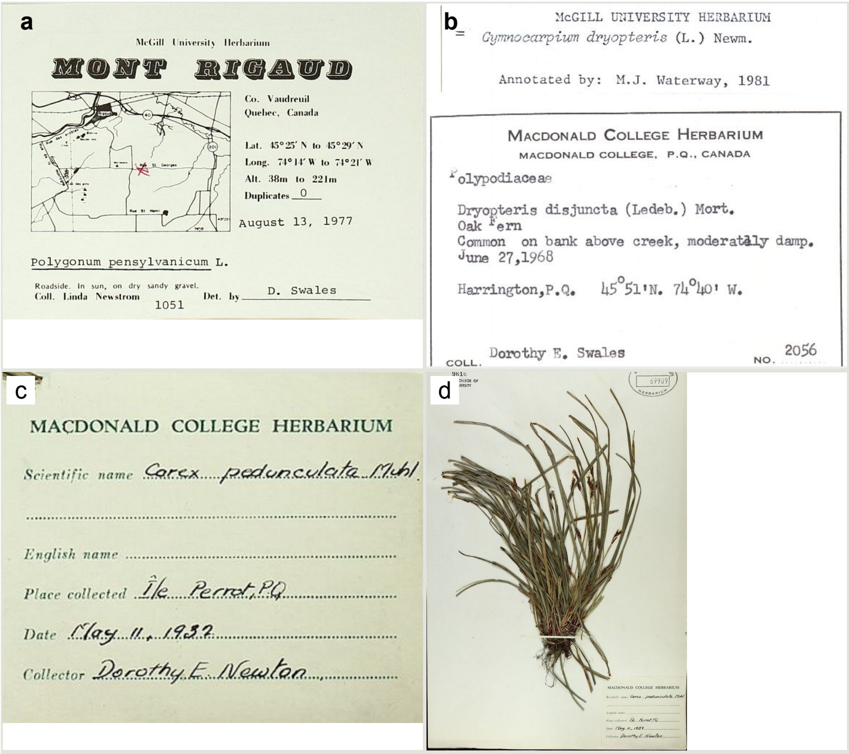
Figure 9. doi

The University of Michigan Class of 1903 Women's Basketball Team with Ethel Winifred Chase, third from the left. It is rare to find pictures of notable women who collected specimens; if they do exist, they are rarely in the portrait style of eminent male collectors. This is the only picture we know of depicting Chase. From the University of Michigan, public domain, via Wikimedia Commons.

Various resources were used to research Chase. They include the Harvard Index of Botanists which contained two entries, a JSTOR Global Plants person database entry, the genealogical research website FamilySearch, which provided her exact birth and death date and a full text search of the Biodiversity Heritage Library corpus, which led to the discovery of a scholarly article on Chase (Jones 1966) . This article outlined the existence of historically significant lantern slides made by Chase during the South Pacific expedition. An email to the University of Michigan Herbarium brought these slides to the attention of curators. The institutional knowledge about these slides had lapsed, but as a result of the disambiguation research being shared with the University of Michigan herbarium staff, historical context as well as attribution data were added to the lantern slide collection. A Wikipedia article on Chase was also created collating the knowledge gained during the disambiguation process.

Dr. Dorothy Swales

Dr. Dorothy Swales was a Canadian botanist and the first female curator of the Macdonald College Herbarium (now known as the McGill University Herbarium). Born in Quebec in 1901, Dr. Swales would attend Macdonald College (later part of McGill University) to earn both undergraduate and graduate degrees in Plant Pathology and Bacteriology, respectively. She would later earn her PhD in Mycology from the University of Manitoba. During her tenure, from 1964 to 1971, Dr. Swales collected extensively throughout Quebec and the Northwest Territories with a specific focus on plants found in the Arctic and sub-Arctic regions (e.g. Fig. 10). Under Dr. Swales, the Herbarium's collection expanded extensively and was enriched through international exchanges of botanical specimens from herbaria in the U.S.S.R., Sweden and Denmark. After retiring in 1971, Dr. Swales continued to engage with the botanical science community at McGill University as emeritus curator.



McGill University Herbarium
MONT RIGAUD
Co. Vaudreuil
Quebec, Canada
Lat. 45°25′ N to 45°29′ N
Long. 74°14′ W to 74°21′ W
Alt. 38m to 221m
Duplicates 0
August 13, 1977
Polygonum pensylvanicum L.
Roadside. In sun, on dry sandy gravel.
Coll. Linda Newstrom 1051
Det. by D. Swales

McGILL UNIVERSITY HERBARIUM
Gymnocarpium dryopteris (L.) Newm.
Annotated by: M.J. Waterway, 1981

MACDONALD COLLEGE HERBARIUM
MACDONALD COLLEGE, P.Q., CANADA
Polypodiaceae
Dryopteris disjuncta (Ledeb.) Mort.
Oak Fern
Common on bank above creek, moderatly damp.
June 27,1968
Harrington,P.Q. 45°51'N. 74°40' W.
COLL. Dorothy E. Swales NO. 2056

MACDONALD COLLEGE HERBARIUM
Scientific name Carex pedunculata Muhl
English name
Place collected Île Perrot, P.Q.
Date May 11, 1932
Collector Dorothy E. Newton

Figure 10.

Labels of specimens of Dr. Dorothy Swales from McGill University Herbarium (Waterway et al. 2018).

a: A specimen determined by Dorothy Swales as "D. Swales" from McGill University Herbarium, Catalogue number 107334 doi
b: The label of a specimen collected by Dorothy Swales as "Dorothy E. Swales" from McGill University Herbarium, Catalogue number 65086. doi
c: A label of a specimen collected by Dorothy Swales as "Dorothy E. Newton" from McGill University Herbarium, Catalogue number 69909. doi
d: The whole specimen collected by Dorothy Swales as "Dorothy E. Newton" from McGill University Herbarium, Catalogue number 69909. doi

Despite Dr. Swales' contributions to the field, both as a botanist and herbarium curator, the lack of a significant or unified digital presence made piecing together the story of her botanical collections and collaborations difficult. Her correspondence, notes and specimens are currently housed at the McGill University Herbarium, but many items in the collection are yet to be digitised. Both professionally and across botanical specimen sheets, Dr. Swales was listed using different variations of her name such as Dorothy E. Swales, Swales Dorothy E, D.E. Swales, Mrs. W.E. Swales (her husband was Dr. William Swales) and Dorothy Newton (her maiden name). Although a search using a variation of her name in individual databases, such as GBIF or Canadensys, might return a result of digitised specimens attributed to Dr. Swales, the different versions of her names, as well as the absence of links between her collections housed in different institutions and available across digital platforms, creates a problem for telling a fuller story of her career as curator and botanist.

The disambiguation process first required the creation of a Wikidata profile and Q number and then the creation of a digital profile on Bionomia. Dr. Swales has now been unambiguously linked to specimens (either as collector or determiner) across 15 organisations. This work has collated her contributions to the McGill Herbarium during her tenure and drawn her collections under one Bionomia profile. Further information was found from her obituary, Google Scholar and resources on McGill University's history relating to the Herbarium and Macdonald College. Unified attribution for Dr. Swales enables a more detailed and clearer narrative of who she was as a botanist, curator and educator. Broadly speaking, as more archival documents (e.g. curator correspondence, field notes) are digitised, the solid establishment of a digital presence will make it easier to add supplemental material information about botanists and their collections, thereby enriching the information to be used for researchers. This will be especially helpful for those focused on the history of women in botanical science.

The next steps for collections

The effort to disambiguate people's names should decrease over time. In fact, it is part of the evolution of collection data management that ends when people are identified as unambiguously as possible. Full disambiguation is many years off, but rapid progress can be made for the vast majority of cases as outlined by Groom et al. (2020) and Güntsch et al. (2021). Below, we propose a number of specific objectives for the next phase of activity. These objectives, which are presented using the SMART framework (Specific, Measurable, Achievable, Relevant, Timely), are written as a direct challenge to collections and their funders to motivate this critical work.

Objective 1: Promote the use of person identifiers in local, national or international outreach, publishing and research activities

This is measurable through the number of person identifiers used in publications and other research outputs. The objective is achievable because the potential for using disambiguated person data (particularly historic data) in scientometrics and biodiversity informatics has not yet been fully realised. This is timely because aggregated person data help us to answer new questions about the relevance of collections, their scientific output and their sociopolitical histories, in addition to supporting policy.

Objective 2: Increase the number of collection management systems that use person identifiers

Modifying data management systems to accommodate person identifiers is relatively simple for most systems, though more sophisticated use of indicators for matching, merging and comparing data is more demanding. Improving software systems is achievable and measurable because software systems are constantly evolving and new ones emerge regularly. Building-in person disambiguation functionality at the design stage is the best strategy. It is timely because collections are increasingly requiring more clarity on person data and software is needed to close the roundtripping cycle.

Objective 3: Increase the number of living collectors registered and using an ORCID identifier when contributing to collections

The uptake of ORCID identifiers can be measured internally by institutions, but also by their use in publications linked to collections and in Wikidata and GBIF. This can be achieved through institutional policies, such as an acquisition policy or data management plan and through promotion of ORCID to collectors who may not currently appreciate how it benefits them. It is relevant because institutions are increasingly being compelled to better manage issues, such as data protection, data sharing and benefit sharing. The advent of GDPR has raised awareness of our rights and responsibilities regarding data on people; increasing the use of ORCID identifiers provides a timely mechanism for better managing person data in-line with GDPR.

Objective 4: Undertake disambiguation in the national languages of many countries

This can be measured by the number of languages for which software, data and training materials are available and can be achieved because Wikidata and Bionomia are already multilingual systems. It is relevant because the disambiguation of names is particularly important for people whose languages are not in Latin scripts and because providing disambiguation guidelines and resources in languages other than English would significantly support adoption. It is also timely because, in the spirit of the Convention on Biological Diversity, the institutions in the Global North have a responsibility to support those countries in the Global South from where many specimens have been obtained.

Objective 5: Increase the number of identified people on Wikidata linked to collections

This is measurable through counts of people and their links on Wikidata, particularly those identified as biologists. It can be achieved through training, community events and projects dedicated to using the results. It is relevant and timely because collections acknowledge their responsibility to recognise the diversity of people who contribute to them and because the tools, specimens and biographical resources are increasingly available digitally online.

Objective 6: Increase the number of people in collections with expertise in person disambiguation

This is measurable through the number of people attending training events on disambiguation of collections and the amount of disambiguation being done. It can be achieved through in-person and online training events, particularly coupled to collections management and informatics conferences, such as those of the Society for the Preservation of Natural History Collections (SPNHC) and the Biodiversity Information Standards (TDWG) organisation. It is relevant because it will enable collections staff to better manage biographical data in their collection and it is timely because it is increasingly easy to disambiguate people and a concerted effort between collections will help the whole collections community.

Objective 7: Collaborate towards an exchange standard for attribution data

To roundtrip person data effectively, a data exchange standard is needed, together with tools for data managers to facilitate decisions about what to confidently accept and what to reject on the return trip. This exchange standard should include data on the source from which new assertions were derived, when they were made, who made them and, ideally, what corroborating evidence was used. It is achievable thanks to the existing W3C Web Annotation Data Model, a model for nanopublications and the report on attribution written by a joint working group of the Research Data Alliance (RDA) and TDWG organisations, which provide a timely foundation for standards development (Thessen et al. 2019).

Conclusions

The current informatics landscape for the disambiguation of people makes it possible to imagine a future where the whole of a person's scientific output is connected. The tools and infrastructure exist to enable and democratise disambiguation of people in collections and there is a clear need. Unlike some other areas of biodiversity informatics, person name disambiguation is an action to which all organisations can contribute and on which lasting and impactful progress can be made. As collections are further digitised, disambiguation will continue, motivated by all the benefits outlined above. We recognise that more work is still needed to disseminate the model for how to do this work, how to share and use these data and how to update current standards of practice that include these identifiers from the beginning. The more people who are disambiguated, the easier the process becomes and the more benefits accrue. While it is likely that tools, databases and collections will change, the broad coalition engaged in disambiguation globally means that there is no single point of failure and we see a bright, interlinked future for collections in which the identities of people will play a pivotal role.

Acknowledgements

The authors would like to thank all the institutions and people who have contributed, often voluntarily, to the disambiguation of people in collections. We thank the reviewers of this paper for their detailed and thoughtful suggestions as they have improved the paper.

We are also grateful to Sven Bellanger for his work on the illustrations.

This paper is a product of the People in Biodiversity Data task group of the Biodiversity Information Standards (TDWG) organisation.

Funding program

This work was supported by European Cooperation in Science and Technology (COST) as part of the Mobilise Action CA17106 on Mobilising Data, Experts and Policies in Scientific Collections; SYNTHESYS+ a Research and Innovation Action (Grant agreement 823827) and DiSSCo Prepare a Coordination and Support Action (Grant Agreement 871043), both funded by the Horizon 2020 Framework Programme of the European Union. This work was also facilitated by the Research Foundation – Flanders (FWO) research infrastructure under grant number I001721N. Additional support provided by the National Science Foundation (NSF) grant number #2033973.

Conflicts of interest

Bionomia is a project developed and maintained by David P. Shorthouse. It does not form part of his official duties as Biodiversity Data Manager with Agriculture and Agri-Food Canada.

References

- Access to Biological Collection Data task group (2019) ABCD 3.0. https://abcd.tdwg.org/3.0/. Accessed on: 2022-4-29.
- Agenjo-Bullón X, Hernández-Carrascal F (2020) Wikipedia, Wikidata y Mix'n'match. Anuario ThinkEPI 14 https://doi.org/10.3145/thinkepi.2020.e14f01
- Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. 3. Proceedings of the International AAAI Conference on Web and Social Media. 361-362 pp. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/13937

- Bhalla N (2019) Strategies to improve equity in faculty hiring. Molecular Biology of the Cell 30 (22): 2744-2749. https://doi.org/10.1091/mbc.e19-08-0476
- Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008 (10). https://doi.org/10.1088/1742-5468/2008/10/p10008
- Bräuchler C, McEwan D, Rainer H (2021) Referencing people in biodiversity research needs persistent unique identifiers and primary data – the example of G.W.H. Schimper. Annalen des Naturhistorischen Museums in Wien. Serie B Für Botanik Und Zoologie 123: 353-366. URL: https://www.jstor.org/stable/26993252
- Constant N, Roberts L (2017) Narratives as a mode of research evaluation in citizen science : understanding broader science communication impacts. Journal of science communication 16 (4): A03. https://doi.org/10.22323/2.16040203
- Copas K (2019) The GDPR-compliance waltz: balancing necessity, legitimate interests and data-subject rights in the GBIF network. Open Science Framework https://doi.org/10.17605/osf.io/9r26b
- European Data Protection Supervisor (EDPS) (2020) A Preliminary Opinion on data protection and scientific research. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf. Accessed on: 2012-4-10.
- Finch J (2008) Naming names: Kinship, individuality and personal names. Sociology 42 (4): 709-725. https://doi.org/10.1177/0038038508091624
- George S (2006) Cultivating the botanical woman: Rousseau, Wakefield and the instruction of ladies in botany. Zeitschrift fur Padagogische Historiographie 12 (1): 3-11. URL: http://hdl.handle.net/2299/2249
- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E (2020) People are essential to linking biodiversity data. Database 2020 https://doi.org/10.1093/database/baaa072
- Groom QJ, O'Reilly C, Humphrey T (2014) Herbarium specimens reveal the exchange network of British and Irish botanists, 1856–1932. New Journal of Botany 4 (2): 95-103. https://doi.org/10.1179/2042349714y.0000000041
- Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, Röpert D, Fichtmüller D, Shorthouse DP, Hyam R, Dillen M, Trekels M, Haston E, Rainer H (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLOS One 16 (12). https://doi.org/10.1371/journal.pone.0261130
- Haak L, Fenner M, Paglione L, Pentz E, Ratner H (2012) ORCID: a system to uniquely identify researchers. Learned Publishing 25 (4): 259-264. https://doi.org/10.1087/20120404
- Hill K (2016) In Search of Useful Collection Metadata: Using OpenRefine to Create Accurate, Complete, and Clean Title-level Collection Information. Serials Review 42 (3): 222-228. https://doi.org/10.1080/00987913.2016.1214529
- Holetschek J, Dröge G, Güntsch A, Berendsohn WG (2012) The ABCD of primary biodiversity data access, Plant Biosystems. An International Journal Dealing with all Aspects of Plant Biology 146 (4): 771-779,. https://doi.org/10.1080/11263504.2012.740085
- Hu Y (2005) Efficient, high-quality force-directed graph drawing. Mathematica journal 10 (1): 37-71.

- International Council of Museums (2017) ICOM Code of Ethics for Museums. International Council of Museums (ICOM), 44 pp. URL: https://icom.museum/wp-content/uploads/2018/07/ICOM-code-En-web.pdf [ISBN 978-92-9012-420-7]
- International Organization for Standardization (2017) Information and documentation — International standard name identifier (ISNI) (ISO 27729:2012). ISNI International Agency. URL: https://www.iso.org/standard/44292.html
- Ishida R (2011) Personal names around the world. https://www.w3.org/International/questions/qa-personal-names. Accessed on: 2022-4-03.
- Jones K (1966) Winifred Chase, intrepid spirit. The Michigan Botanist 5 (4): 183-191. URL: https://www.biodiversitylibrary.org/page/57502930
- Larsson A, Ånäs S, Zeinstra M (2019) Wikimedia Commons Data Roundtripping, Final Report. URL: https://upload.wikimedia.org/wikipedia/commons/e/e8/Wikimedia_Commons_Data_Roundtripping_-_Final_report.pdf
- Lewis G (2004) The role of museums and the professional code of ethics. In: Boylan P (Ed.) Running a museum: A practical handbook. ICOM – International Council of Museums, Paris, 1-16 pp. URL: https://www.museumsverband-rlp.de/fileadmin/user_upload/Verbandsseite/running_a_museum.pdf [ISBN 92-9012-157-2].
- Lindon H, Gardiner L, Brady A, Vorontsova M (2015) Fewer than three percent of land plant species named by women: Author gender over 260 years. Taxon 64 (2): 209-215. https://doi.org/10.12705/642.4
- Macdonald R (2018) A code of ethics for United States museums. Museum International 45 (1): 53-56. https://doi.org/10.1111/j.1468-0033.1993.tb01092.x
- Maroske S, Vaughan A (2014) Ferdinand Mueller's female plant collectors: a biographical register. Muelleria 32: 92-172. https://doi.org/10.5962/p.295690
- Mast A, Paul D, Rios N, Bruhn R, Dalton T, Krimmel E, Pearson K, Sherman A, Shorthouse D, Simmons N, Soltis P, Upham N, Abibou D (2021) Rapid Creation of a Data Product for the World's Specimens of Horseshoe Bats and Relatives, a Known Reservoir for Coronaviruses (1.6). Zenodo https://doi.org/10.5281/zenodo.5044247
- Meeker N, Szabari A (2020) Radical botany: Plants and speculative fiction. Fordham University Press [ISBN 9780823286638] https://doi.org/10.5422/fordham/9780823286638.001.0001
- Meeus S, August T, Trekels M, Reyserhove L, Groom QJ (2021) Network analysis of specimen co-collection. BioHackrXiv https://doi.org/10.37044/osf.io/4ahng
- Nicolson N, Paton A, Phillips S, Tucker A (2018) Specimens as Research Objects: Reconciliation Across Distributed Repositories to Enable Metadata Propagation. 2018 IEEE 14th International Conference on e-Science (e-Science) https://doi.org/10.1109/escience.2018.00028
- Osborn D, Helmy I (1980) The contemporary land mammals of Egypt (including Sinai) / Dale J. Osborn, Ibrahim Helmy. Fieldiana., Zoology (5). https://doi.org/10.5962/bhl.title.2801
- Page R (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e8767
- Rabeler R, Svoboda H, Thiers B, Prather LA, Macklin J, Lagomarsino L, Majure L, Ferguson C (2019) Herbarium Practices and Ethics, III. Systematic Botany 44 (1): 7-13. https://doi.org/10.1600/036364419x697840

- Shorthouse D (2020) Slinging With Four Giants on a Quest to Credit Natural Historians for our Museums and Collections. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59167
- Shteir A (1996) Cultivating women, cultivating science: Flora's daughters and botany in England, 1760-1860. Johns Hopkins University Press, London, 301 pp. [ISBN 0-8018-5141-6]
- Smalheiser N, Torvik V (2011) Author name disambiguation. Annual Review of Information Science and Technology 43 (1): 1-43. https://doi.org/10.1002/aris.2009.1440430113
- Staunton C, Slokenberga S, Mascalzoni D (2019) The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks. European Journal of Human Genetics 27 (8): 1159-1167. https://doi.org/10.1038/s41431-019-0386-5
- Sterner E (2019) Cleaning collections data using OpenRefine. Issues in Science and Technology Librarianship 92 https://doi.org/10.29173/istl30
- Thessen A, Woodburn M, Koureas D, Paul D, Conlon M, Shorthouse D, 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). https://doi.org/10.5334/dsj-2019-054
- Triebel D, Scholz P (2022) IndExs – Index of Exsiccatae. Botanische Staatssammlung München. Release date: 2001-2-01. URL: http://indexs.botanischestaatssammlung.de
- Valdecasas A (2011) An index to evaluate the quality of taxonomic publications. Zootaxa 2925 (1). https://doi.org/10.11646/zootaxa.2925.1.6
- van Hooland S, De Wilde M, Verborgh R, Steiner T, Van de Walle R (2013) Exploring entity recognition and disambiguation for cultural heritage collections. Digital Scholarship in the Humanities 30 (2): 262-279. https://doi.org/10.1093/llc/fqt067
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602
- Waterway M, Martins K, Larocque G (2018) McGill University Herbarium Database. 2.1. McGill University. Release date: 2018-12-20. URL: https://doi.org/10.5886/srzbj7
- Weil S (2007) Beyond big and awesome: Outcome-based evaluation. In: Sandell R, Janes R (Eds) Museum Management and Marketing. Routledge, London, 448 pp. https://doi.org/10.4324/9780203964194
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS One 7 (1). https://doi.org/10.1371/journal.pone.0029715
- Woolston C (2022) Time to rethink the scientific CV. Nature 604 (7904): 203-205. https://doi.org/10.1038/d41586-022-00928-4

Supplementary materials

Suppl. material 1: A disambiguation strategy doi

Authors: Elspeth M. Haston, Christian Bräuchler, Robert WN Cubey, Mathias Dillen, Pieter Huybrechts, Nicole Kearney, Niels Klazenga, Siobhan Leachman, Deborah L. Paul, Heather Rogers, Joaquim Santos, David Peter Shorthouse, Alison Vaughan, Sabine von Mering, Quentin Groom

Data type: flow diagram

Brief description: A diagrammatic representation of one disambiguation strategy. Strategies vary considerably depending on the name being disambiguated, the dates involved, the taxonomy of the specimen, the collecting locality and the collection it is held in. Abbreviations: Harvard University Herbaria Index of Botanists (HUH); International Plant Names Index (IPNI); Royal Botanic Garden Edinburgh (RBGE).

Download file (536.70 kb)

Suppl. material 2: Disambiguation flow example doi

Authors: Elspeth M. Haston, Christian Bräuchler, Robert WN Cubey, Mathias Dillen, Pieter Huybrechts, Nicole Kearney, Niels Klazenga, Siobhan Leachman, Deborah L. Paul, Heather Rogers, Joaquim Santos, David Peter Shorthouse, Alison Vaughan, Sabine von Mering, Quentin Groom

Data type: flow diagram

Brief description: A real example of how a name string is disambiguated and the steps taken in documenting it.

Download file (548.98 kb)