

People in the biodiversity knowledge graph and their roles in building the data connections we need

26 July 2023

Erica Krimmel & Holly Little

BOTANY 2023, Special Lecture: Supporting inclusive and sustainable research infrastructure for systematics (SISRIS) by connecting scientists and their specimens

Abstract: Herbaria are connected to each other by intricate and deep-rooted social histories. A student trained at one collection moves on and eventually becomes the curator of another. A prolific amateur collector donates specimens to multiple herbaria over the course of their lifetime. Rival systematists wage a decades-long battle documented by annotations back-and-forth on specimen labels. Although 21st-century data management in herbaria has not prioritized information about the people associated with specimens, people are often a critical link to data beyond the basic specimen occurrence record. Capturing and sharing more data about the “who” of specimens can improve connections across institutions and individuals, augment local data records, and encourage expertise-sharing. Typically, data about people are digitized and managed individually by each herbarium or institution, or at best by a consortia of institutions using the same collections management system. Not only does this lead to redundant time spent, but it also results in isolated knowledge management. Shared knowledge management, in contrast, can improve knowledge completeness, raise the visibility of the work required to manage knowledge, and make data more accessible to the linked open data ecosphere. Ultimately, these benefits lead to improved discoverability for specimens by increasing their data connectivity in the biodiversity knowledge graph.

Over the past few years, Wikidata has gained visibility in the biodiversity collections community as a centralized, accessible platform for working collaboratively to disambiguate entities, e.g., people associated with herbaria, and to mobilize information about them. In this way, Wikidata is a tool for shared knowledge management, and we can use it to support inclusive and sustainable research infrastructure. Such research infrastructure depends on social systems as much as on

information systems for successful knowledge management. Wikidata also provides an established social system with its collaborative, community-oriented approach to curation. This approach may be initially uncomfortable to many herbarium professionals, but relinquishing total control over “our” data promotes inclusivity by recognizing that we may not be the ultimate authorities on every aspect of our collections data. This is especially true of data related to people, who are frequently important to domains other than herbaria. Even within the herbarium community, many individuals involved with collections are not fully acknowledged for their work or have been misrepresented, especially those who are women, non-White, and/or Indigenous. Tools like Wikidata offer the opportunity for data to be augmented and/or corrected, and for this work to be done in a shared knowledge management context that benefits all herbaria and specimens connected to an individual.

People connect collections...



Key talking points: (1:15)

- Herbaria are connected to each other by intricate and deep-rooted social histories. A student trained at one collection moves on and eventually becomes the curator of another. A prolific amateur collector donates specimens to multiple herbaria over the course of their lifetime. Rival systematists wage a decades-long battle documented by annotations back-and-forth on specimen labels.
- Although 21st-century data management in herbaria has not prioritized information about the people associated with specimens, people are often a critical link to data beyond the basic specimen occurrence record.
- Capturing and sharing more data about the “who” of specimens can improve connections across institutions and individuals, augment local data records, and encourage expertise-sharing.
- In this talk, using Elizabeth Atwater as a throughline. Atwater was an amateur botanist who lived in the Midwest in the mid-1800s and collected plants both near her home and on travels to the western US. Shown on this slide are institutions that hold specimens collected or identified by Atwater.

NOTES:

- Elizabeth Atwater on Bionomia: <https://bionomia.net/Q66581882>
- Image source: https://commons.wikimedia.org/wiki/File:Elizabeth_Emerson_Atwater.jpg

...but collections aren't connecting people



Key talking points: (1:10)

- This slide shows several of the many name variations used for Elizabeth Atwater on her specimens and in the digitized specimen data available on aggregators such as GBIF. Disambiguating these name variations can be easy or difficult, depending on the context. Note that in one of the variations Elizabeth Atwater is only identified by her husband's name, which is very common with female collectors of this era.
- Using text strings to identify a person doesn't lead to strong connections because text can be so variable.
- Typically, data about people are digitized and managed individually by each herbarium or institution, or at best by a consortia of institutions using the same collections management system. Not only does this lead to redundant time spent, but it also results in isolated knowledge management and fails to make connections between the same concept represented in different datasets.

NOTES:

- Image source:
https://commons.wikimedia.org/wiki/File:Elizabeth_Emerson_Atwater.jpg

Identifiers are a tool to connect people

The image displays three specimen labels and a GBIF screenshot. The top-left label is from the Herbarium of The Chicago Academy of Sciences, with handwritten text: 'LOCALITY Potomac, Virginia', 'DATE June 1874', 'COLLECTOR Maria', 'HABITAT flower lavender color', and 'E.E. Atwater colln IDENTIFIED BY'. The middle label is from the Herbarium N. Y. State Museum, with handwritten text: 'Name *Lentiginos Ledebur*', 'Locality Virginia', 'Collected by Mrs. E. E. Atwater', and 'No. Date'. The bottom-left label is from the Coll. N. Y. State, with handwritten text: 'Non *Pringicula lutea Walt.*', 'Loc. Jacksonville, Fla.', 'Leg. Mrs. E. A. Atwater', and 'Mb70-F3-3000'. The top-right screenshot shows a 'Recorded by' dropdown menu with three checked entries: 'Atwater, Elizabeth E.', 'Atwater, E. E.', and 'Collector(s): Elizabeth Emerson Atwater'. The bottom-right screenshot shows a search for 'Occurrences' with '0 RESULTS' and a table with columns 'ID', 'METRICS', and 'DOWNLOAD'. The table contains two rows: one with ID 'http://www.wikidata.org/entity/Q66581882' and another with ID 'Q66581882'. The 'Recorded by ID' dropdown is set to 'Recorded by ID'.

Key talking points: (1:20)

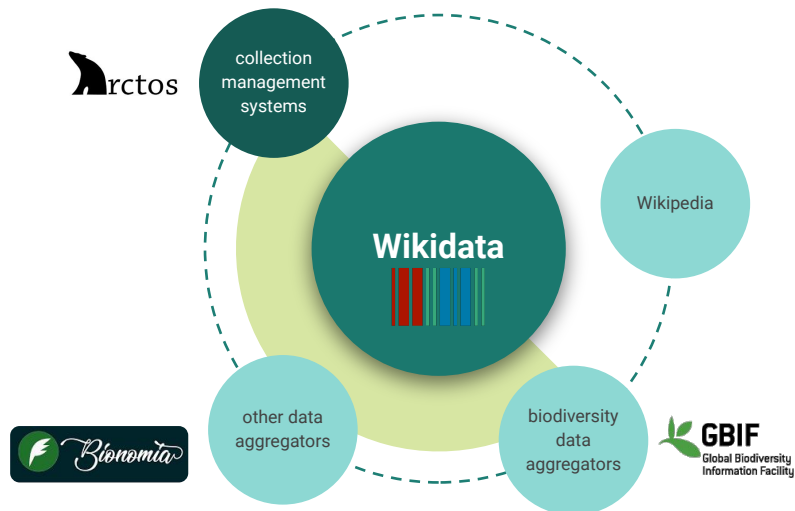
- The idea of connecting the same concept represented in different datasets is crucial to the biodiversity knowledge graph, which itself is an idea that we can contribute massive amounts of specimen (and other) data into a network that is interpretable by both humans and computers.
- In a biodiversity knowledge graph, people can be valuable entry points into aggregated data, but only if they are discoverable.
- Rather than using text strings, we can use unique identifiers to make explicit connections between the same person in different datasets.
- We aren't doing this very well yet, as shown in this slide. The specimen labels on the left all have different variations of Atwater's name, and the GBIF screenshot on the top right shows how the collector from each of these labels was transcribed into the local herbarium's database. None of these text strings even match the text as written on the label, much less each other.
- In the screenshot on the bottom right, we see that searching GBIF specimen data by an identifier used for Atwater gets us... zero results. Ideally, each of the individual specimen records would record Atwater's identifier in their local data and then it wouldn't matter how they transcribed her name because all name variations would be connected to this unique identifier.

NOTES:

- Labels from <https://www.gbif.org/occurrence/3413423261>,
<https://www.gbif.org/occurrence/473431739>,

- <https://www.gbif.org/occurrence/3414040169>

Wikidata is a tool to use identifiers



Key talking points: (1:20)

- Identifiers are easier to talk about than to implement, which is why we haven't solved this problem already.
- Over the past few years, Wikidata has gained visibility in the biodiversity collections community as a tool for using identifiers, and a complement to other cyberinfrastructure components, like collection management systems and data aggregators.
- Wikidata is a centralized, accessible platform for working collaboratively to disambiguate entities, e.g., people associated with herbaria, and to mobilize information about them. In this way, Wikidata is a tool for shared knowledge management, and we can use it to support inclusive and sustainable research infrastructure.
- Shared knowledge management can improve knowledge completeness, raise the visibility of the work required to manage knowledge, and make data more accessible to the linked open data ecosphere. Ultimately, these benefits lead to improved discoverability for specimens by increasing their data connectivity in the biodiversity knowledge graph.
- In the next few slides we'll explore how Elizabeth Atwater is or could be connected between Wikidata, Arctos (a collection management system), Bionomia, and GBIF.

NOTES:

-



Identifiers enable shared knowledge management



Name Variations for Elizabeth Emerson Atwater

- Elizabeth (first name) [search]
- Emerson (middle name) [search]
- Atwater (last name) [search]
- Elizabeth Atwater (aka) [search]
- Elizabeth E. Atwater (aka) [search]
- Mrs. Samuel T. Atwater (aka) [search]

<https://arctos.database.museum/agent/21295246>

Language	Label	Description	Also known as
English	Elizabeth Emerson Atwater	naturalist, botanist and botanical collector (1812-1878)	E. E. Atwater Ella Emerson Ella Atwater Elizabeth Atwater
Spanish	Elizabeth Emerson Atwater	No description defined	

<https://www.wikidata.org/wiki/Q66581882>


Key talking points: (0:25)

- The name variations issue we started this talk with encapsulates a major benefit of using identifiers to share the burden of knowledge management.
- In this example, both Arctos and Wikidata track variations of Atwater's name. They are not coordinated and thus have different information.
- Identifiers provide a mechanism by which different cyberinfrastructure components could exchange information and, ideally, synchronize what they each know.

NOTES:

- <https://arctos.database.museum/agent/21295246>
- <https://bionomia.net/Q66581882>
- <https://www.wikidata.org/wiki/Q66581882>


Identifiers enable shared knowledge management



Overview Specialties Network Deposited At

Co-collectors Identified For Identifications By

Has collected with:




Floretta Allen Curtiss
 * December 01, 1822 – March 03, 1899 †
 United States

Collected Rhodomelaceae and identified Halimedaceae

415 specimens claimed DOI 10.5281/zenodo.8034076

<https://bionomia.net/Q66581882>



Relationships To Elizabeth Emerson Atwater

From	Relationship	Begin	End	Remark
Samuel Tyler Atwater	spouse of			
Increase Allen Lapham	associate of			
Emma M. Taylor	associate of			
Mrs. E. Aiken	associate of			
Mrs. William Gooding	associate of			
E. P. Stevens	associate of			
Mrs. Gale	associate of			
Mrs. Hilliard	associate of			
Mrs. Pearson	associate of			
Mrs. George P. Hanson	associate of			
Mrs. Patrick Anderson	associate of			Travelled with Elizabeth Emerson Atwater to the United Kingdom, France, and Italy [weaver: Did Atwater travel outside of the US?]
Mrs. George Vail	associate of			

<https://arctos.database.museum/agent/21295246>

Key talking points: (0:45)

- We're all familiar with networks of people. People who collect together, whose research impacts each other, who are friends or spouses or colleagues with each other. In the biodiversity knowledge graph, understanding the connections between people can provide context for specimens.
- Both Bionomia and Arctos make an effort to capture connections between people. For Atwater, the connections that Bionomia and Arctos know about are non-overlapping.
- Again, if these two systems were able to speak to each other, our view of Atwater's human network would be more complete.

NOTES:

- <https://arctos.database.museum/agent/21295246>
- <https://bionomia.net/Q66581882>
- <https://www.wikidata.org/wiki/Q66581882>

Identifiers enable shared knowledge management

Bionomia

Overview Specialties Network Deposited At Specimens Science Enabled

Elizabeth Emerson Atwater

E. E. Atwater; Ella Emerson; Ella Atwater; Elizabeth Atwater
 * August 08, 1812 – April 11, 1878 *
 botanist, botanical collector, naturalist
 naturalist, botanist and botanical collector (1812-1878)

2,220 specimens collected from at least 24 countries
 16 specimens identified from at least 1 country
 31 specimens used in 9 works

Collected From Map List
 Identified From Map List
 Dates Collected Chart List

United States

<https://bionomia.net/Q66581882>



Collection Activity by Elizabeth Emerson Atwater

Role	Collection	RecordCount	Link
(any)	(all)	2417	Open Catalog Record Results
collector	(all)	2417	Open Catalog Record Results
(any)	CHAS:EH	26	Open Catalog Record Results
collector	CHAS:EH	26	Open Catalog Record Results
(any)	CHAS:Ento	1	Open Catalog Record Results
collector	CHAS:Ento	1	Open Catalog Record Results
(any)	CHAS:Herb	2386	Open Catalog Record Results
collector	CHAS:Herb	2386	Open Catalog Record Results
(any)	CHAS:Inv	4	Open Catalog Record Results
collector	CHAS:Inv	4	Open Catalog Record Results

Media Associated with Elizabeth Emerson Atwater Collection Activity

- 961 Media records referencing collected/prepared catalog records

<https://arctos.database.museum/agent/21295246>

Key talking points: (0:30)

- If we want to explore Atwater's collecting activity, we could do so via many different facets: geography, taxonomy, type of specimen, timespan, etc.
- Arctos and Bionomia have gathered and summarized information for us about Atwater's collecting activity, each focusing on different facets.
- If these two systems were able to speak to each other, we could have richer insight into Atwater's collecting activity.

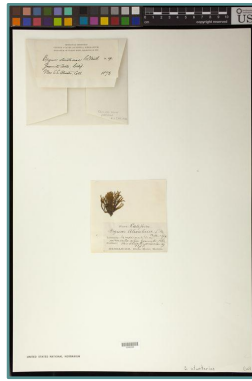
NOTES:

- <https://arctos.database.museum/agent/21295246>
- <https://bionomia.net/Q66581882>
- <https://www.wikidata.org/wiki/Q66581882>

Identifiers enable shared knowledge management



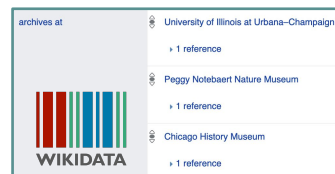
Ramirez J, Watson K, McMillin L, Cjeli E, Sessa E (2023). The New York Botanical Garden Herbarium (NY). Version 1.62. The New York Botanical Garden. Occurrence dataset <https://doi.org/10.15468/6e8nje> accessed via GBIF.org on 2023-07-20. <https://www.gbif.org/occurrence/1930479861>



Orrell T, Informatics Office (2023). NMNH Extant Specimen Records (USNM, US). Version 1.70. National Museum of Natural History, Smithsonian Institution. Occurrence dataset <https://doi.org/10.15468/hnhrg3> accessed via GBIF.org on 2023-07-20. <https://www.gbif.org/occurrence/1456283672>



Kennedy J (2023). Harvard University Herbaria: All Records. Harvard University Herbaria. Occurrence dataset <https://doi.org/10.15468/63pvnth> accessed via GBIF.org on 2023-07-20. <https://www.gbif.org/occurrence/1995082509>



<https://www.wikidata.org/wiki/Q66581882>

Key talking points: (2:00)

- From a specimen perspective—which is the perspective those of us here are probably most familiar with—connecting information about collectors can supplement what we know, particularly for historic specimens.
- For example, Atwater collected a moss in Yosemite in 1873 that was described as a new species, *Bryum atwateriae*. Although the species has since been synonymized, her specimens remain types and are important to taxonomic history. But these specimens were collected 150 years ago and the information present on each of three types is variable.
- In the leftmost image, NYBG has a poor specimen but with lots of historical annotations.
- In the center image, NMNH has a better specimen but very little information and some difficult to read handwriting.
- In the top rightmost screenshot, Harvard hasn't made an image of this specimen available, although they have transcribed the label data.
- Without some very intentional searching, you wouldn't know that these specimens are related. In fact, that they *are* related is only made clear by ancillary information from Atwater's archives. These archives also provide valuable context about the collecting event for these types.
- Using an identifier for Atwater in the specimen data would allow someone to then follow breadcrumbs to find her archives, which are curated separately from any of her specimens, and in part by institutions that don't even have a foothold in the biodiversity science community. You can see in the lower rightmost screenshot that Wikidata has information about where Atwater's

- archives are held.

NOTES:

- <https://arctos.database.museum/agent/21295246>
- <https://bionomia.net/Q66581882>
- <https://www.wikidata.org/wiki/Q66581882>

- <https://w.wiki/75Si>
- <https://w.wiki/75Sn>

Connected data changes collections management

Last: (Perso...	First: (...	Middle: (P...
Gardner	J.	A.
Gardner	Julia	A.
Gardner	Julia	Anna
Gardner	Julie	
Gardner		

Coll... *Est. T.* Ident... *Est. Trees. bnll.*

Collectors: 1 Moore, Ellen James,

GUIDs

1	yes	Type	GUID
*		Wikidata ID	Q84708760

Inclusion of Wikidata ID in EMu

Key talking points:

- Using Wikidata more will have an effect on day-to-day internal collection data practices and institutional data models. For example, harmonizing names and documenting their aliases through shared knowledge management in Wikidata completely changes our ability to manage people data locally.
- In the screenshots here, you can see that even if I have all these variations within my own collection data, I can work to integrate the Wikidata Q #s into our database to help contend with those variations and to assist data entry when translating from one verbatim name on a label to the documented name in the database record.
- Although it may be initially uncomfortable to many herbarium professionals, relinquishing total control over “our” data has the potential to not only lessen the workload on collections staff but also to promote inclusivity by recognizing that we may not be the ultimate authorities on every aspect of our collections data.

NOTES:

- Ida Shepard: <https://www.wikidata.org/wiki/Q21664806>
- NMNH EMu: 4 records for Ellen James Moore, 152 event/site records, 2555 specimen records

Connected data highlights expertise



Key talking points: (1:00)

- This is especially true of data related to people, who are frequently important to domains other than herbaria. Even within the herbarium community, many individuals involved with collections are not fully acknowledged for their work or have been misrepresented, especially those who are women, non-White, and/or Indigenous.
- Although Atwater made valuable contributions to science, she wasn't a prominent figure of her era. Without additional context, herbaria that only have a specimen or two collected by Atwater likely lack the ability to recognize these specimens as originating from a place of expertise.
- Identifiers provide a way to connect people like Atwater into our knowledge of biodiversity.
- Tools like Wikidata offer the opportunity for data to be augmented and/or corrected, and for this work to be done in a shared knowledge management context that benefits all herbaria and specimens connected to an individual.

NOTES:

-



Get started with community-created, adaptable guidelines

Guidelines for Using Wikidata to Mobilize Information about People in Collections: A Paleontology Perspective



<https://doi.org/10.5281/zenodo.6977243>

Created by: (authors listed alphabetically) [Jennifer Bauer](#), [Roger Burkhalter](#), [Talia Karim](#), [Erica Krimmel](#), [Margaret Landis](#), [Siobhan Leachman](#), [Holly Little](#), [Malena Lorente](#), [Suzanne K. Mills](#), [Nicole Neu-Yagle](#), [Ben Norton](#), [Deborah Paul](#), [David Shorthouse](#), [Jessica Utrup](#), [Jacob Van Veldhuizen](#), and [Lindsay Walker](#), with input from participants of the *Using Wikidata to Capture and Share Information about People in Paleontology* workshop.

NOTE Should you wish to credit this document, please cite as: Bauer J, Burkhalter R, Karim T, Krimmel E, Landis M, Leachman S, Little H, Lorente M, Mills SK, Neu-Yagle N, Norton B, Paul D, Shorthouse D, Utrup J, Van Veldhuizen J, and Walker L. 2022. *Guidelines for Using Wikidata to Mobilize Information about People in Collections: A Paleontology Perspective*. <https://doi.org/10.5281/zenodo.6977243>

License: [CC0 1.0 Universal Public Domain Dedication](#)

Introduction

Purpose of this document

The purpose of this document is to provide a framework for how to mobilize information via Wikidata about people working in and/or associated with scientific collections. Building on previous Wikidata documentation produced by

Key talking points: (1:00)

- Our current reality is not one where people are all connected into the biodiversity knowledge graph and we can seamlessly navigate through rich forests of information. However, we aren't far off and the technology to get from here to there largely exists.
- If you are interested in exploring Wikidata, I wanted to share this resource created by the Paleo Data Working Group: "Guidelines for Using Wikidata to Mobilize Information about People in Collections: A Paleontology Perspective." These guidelines are designed to lay out conventions for creating and editing Wikidata items about people connected to biodiversity collections, as well as to serve as a step-by-step learning resource.
- While the examples used in this document all relate to people associated with paleontology, we wrote it to be general enough for the broader community to use, and we also published it in the public domain to encourage maximum uptake and reuse.

NOTES:

Thank you!

Get in touch...



Erica Krimmel

[ORCID 0000-0003-3192-0080](https://orcid.org/0000-0003-3192-0080)

ekrimmel@gmail.com



Holly Little

[ORCID 0000-0001-7909-4166](https://orcid.org/0000-0001-7909-4166)

littleh@si.edu

Key talking points: ()

-

NOTES:

-