Building a Digital Extended Specimen One Association at a Time: What Does It Take to Extend OBI Herbarium Records with their Associated GenBank Sequences?

Jorrit H. Poelen

Ronin Institute / UCSB Cheadle Center for Biodiversity and Ecological Restoration / Global Biotic Interactions

20 Sept 2023 @ BioDigiCon 2023

Guiding Questions

What do *you* do to realize the Digital Extended Specimen [1]?

What would you like *others* to do to realize the Digital Extended Specimen?

[1] Hardisty et al. 2022, BioScience doi:10.1093/biosci/biac060

Guiding Questions (personalized)

What does *Jorrit* do to realize the Digital Extended Specimen?

What would *Jorrit* like *others* to do to realize the Digital Extended Specimen?

What does *Jorrit* do to realize the Digital Extended Specimen?

1. Communicate, Collaborate, and Explore
2. Remix, Reuse Existing Tools/Data
3. (Only If Absolutely Needed) Try Something New
4. Goto 1.

Communicate, Collaborate, and Explore

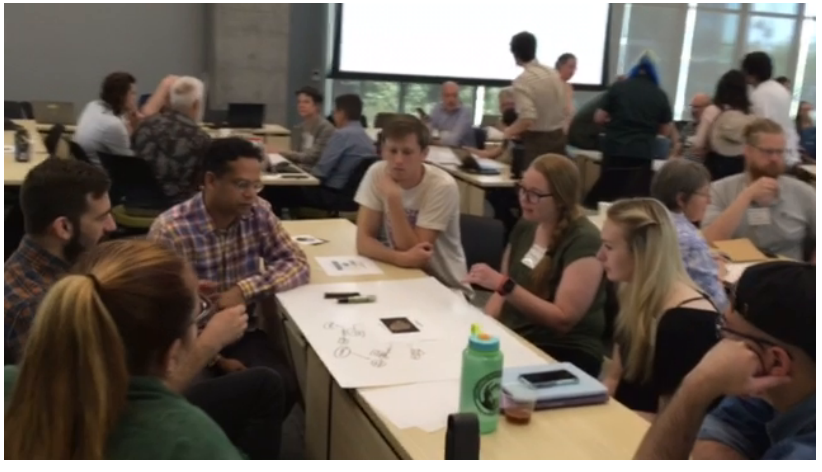
Digital Data In Biodiversity Research 2023 @ ASU

Workshop: Addressing Roadblocks and Envisioning Solutions of the Digital Extended Specimen

The goal of the extended specimen

Welcome

Jenn Yost, Katie Pearson, Lindsay Walker, Jorrit Poelen



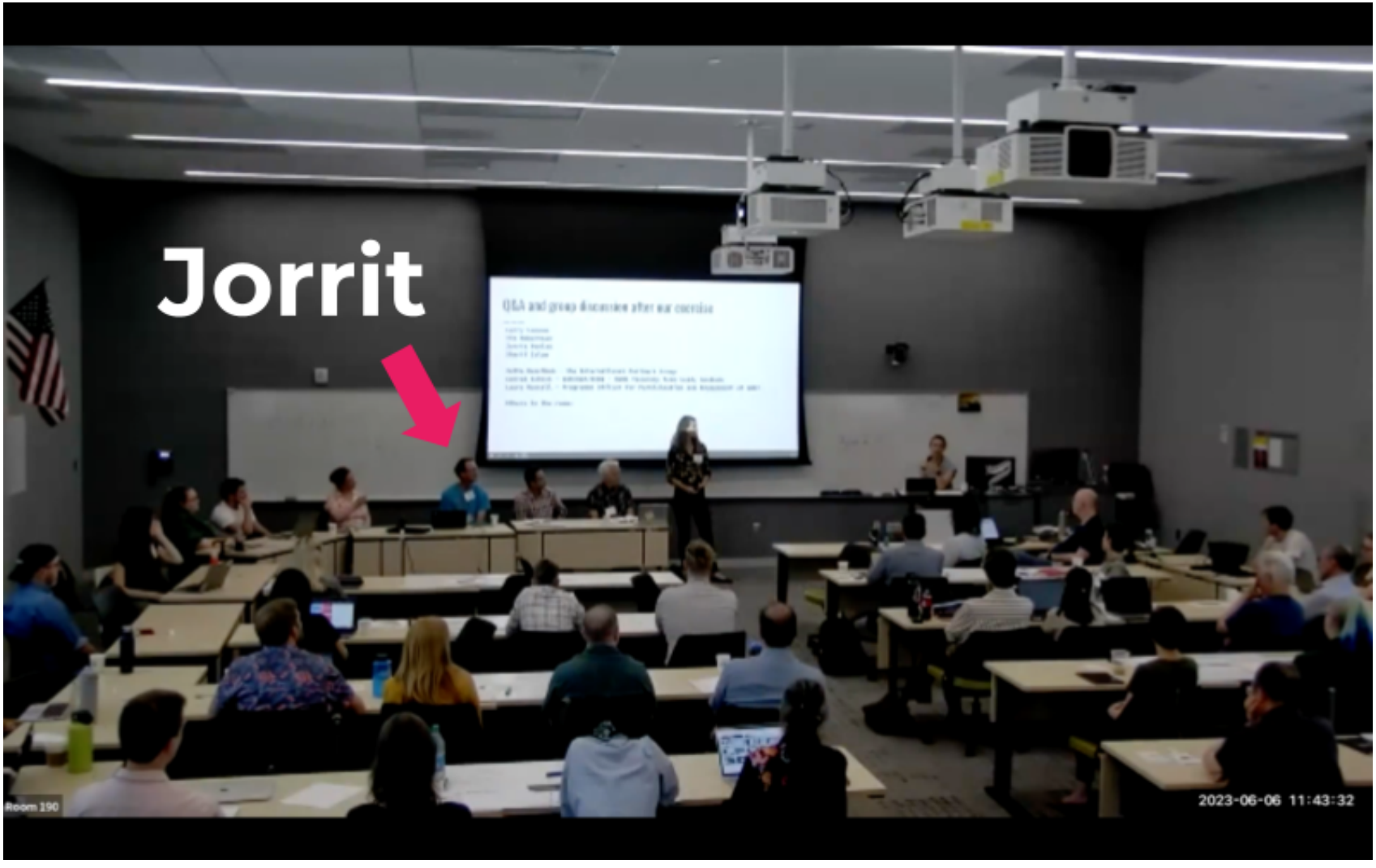
Communicate, Collaborate, and Explore

Communicate, Collaborate, and Explore

> *"[. . .] As a community, the most basic extended specimen is: Here's the specimen and here's its sequence data. That is the link that everbody wants all the time. [. . .]"*

— Jenn Yost 2023. youtu.be/CNRAJvyDHu8?t=9713

Communicate, Collaborate, and Explore



Communicate, Collaborate, and Explore

> *"[. . .] you're looking [. . .] to review your data set and find all the [. . .] sequences known for your specimen [. . .] I'd say let's build it let's do it as long as you give me a coffee and a cookie [. . .]*

— Jorrit Poelen 2023. youtu.be/CNRAJvyDHu8?t=10080

Communicate, Collaborate, and Explore

> *"[. . .] you're looking [. . .] to review your data set and find all the [. . .] sequences known for your specimen [. . .] I'd say let's build it let's do it as long as you give me a coffee and a cookie [. . .]* ***I'm serious.***"

— Jorrit Poelen 2023. youtu.be/CNRAJvyDHu8?t=10080

Extending Specimen: Visualizing the “it” in “let’s build it”

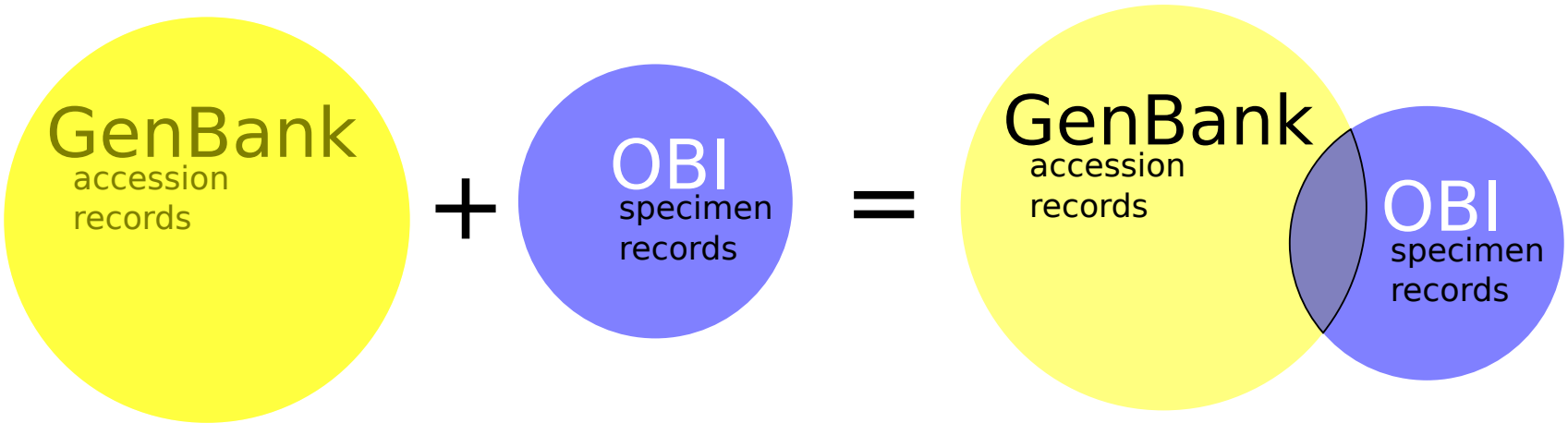


Figure 1: OBI specimens overlap GenBank [2] records.

[2] Sayers et al. 2019. Nucleic Acids Research DOI:10.1093/nar/gky989

Intermezzo

Dunn's ideas in his 1946 "Record Linkage" publication sure do sound a lot like ideas behind the (digital) extended specimen, don't they? [3]

Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service, Federal Security Agency, Washington, D. C.

[3] Dunn HL. 1946. Record Linkage. American Public Health Association. doi:10.2105/AJPH.36.12.1412

Trying Existing Tools

Figure 2: Search for "OBI" in NCBI GenBank Online Search via https://www.ncbi.nlm.nih.gov/nuccore/?term=OBI yielded over 200k candidate records at 2023-09-19.

OBI-GenBank Integration Workflow

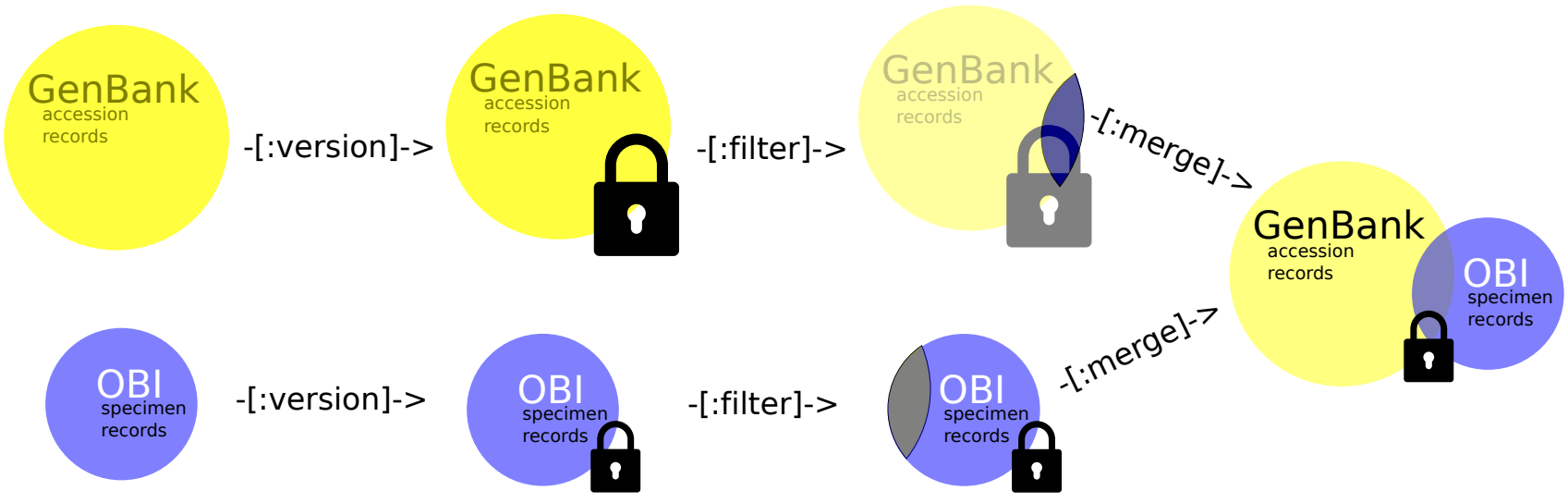


Figure 3: A workflow extending OBI specimen to include associated GenBank records enabled through versioning and data streaming using existing tools and resources (e.g., GenBank, Symbiota, Preston, grep, jq, Zenodo, BioKIC@ASU)

OBI-GenBank Integration Workflow Details

We found, among others, OBI associated GenBank accession **MT735455**, using my Lenovo T480s Laptop, BioKIC, and a 500Mb internet connection by [4]:

1. versioning GenBank PLN [5]
2. streaming all ~7M GenBank PLN records
3. filter by GenBank record containing "OBI", not followed or preceded by letters
4. reviewing the ~200 resulting records
5. verify linked OBI specimen records

---

[4] Poelen JH, Pearson KD, Yost J. 2023. Extending OBI Herbarium Records to include associated NCBI GenBank sequences. https://jhpoelen.nl/obi-genbank hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb

[5] Poelen, Jorrit H. (2023). GenBank PLN (Plantae, Fungi, Algae) Sequence Index in TSV, CSV, JSONL formats hash://sha256/bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce6751fa4b870 hash://md5/f6f78f64e3b3ff06adc3229badbd578b (0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8117720

Example: Herbarium Specimen **OBI09031**

The Hoover Herbarium (OBI) hosts a preserved specimen of type *Angelica hendersonii* Coult. & Rose that was collected in 1966-07-05 by Tracey & Viola Call at the north end of Tomales Bay and 2 mi south of Tomales in Marin County, California with catalog number: **OBI09031**, collector number: 2490, occurrence id: 256368e3-f8d7-4028-8010-1a4ff3eb8111, and web reference https://cch2.org/portal/collections/individual/index.php?occid=166203.

Example: Herbarium Specimen **OBI09031**

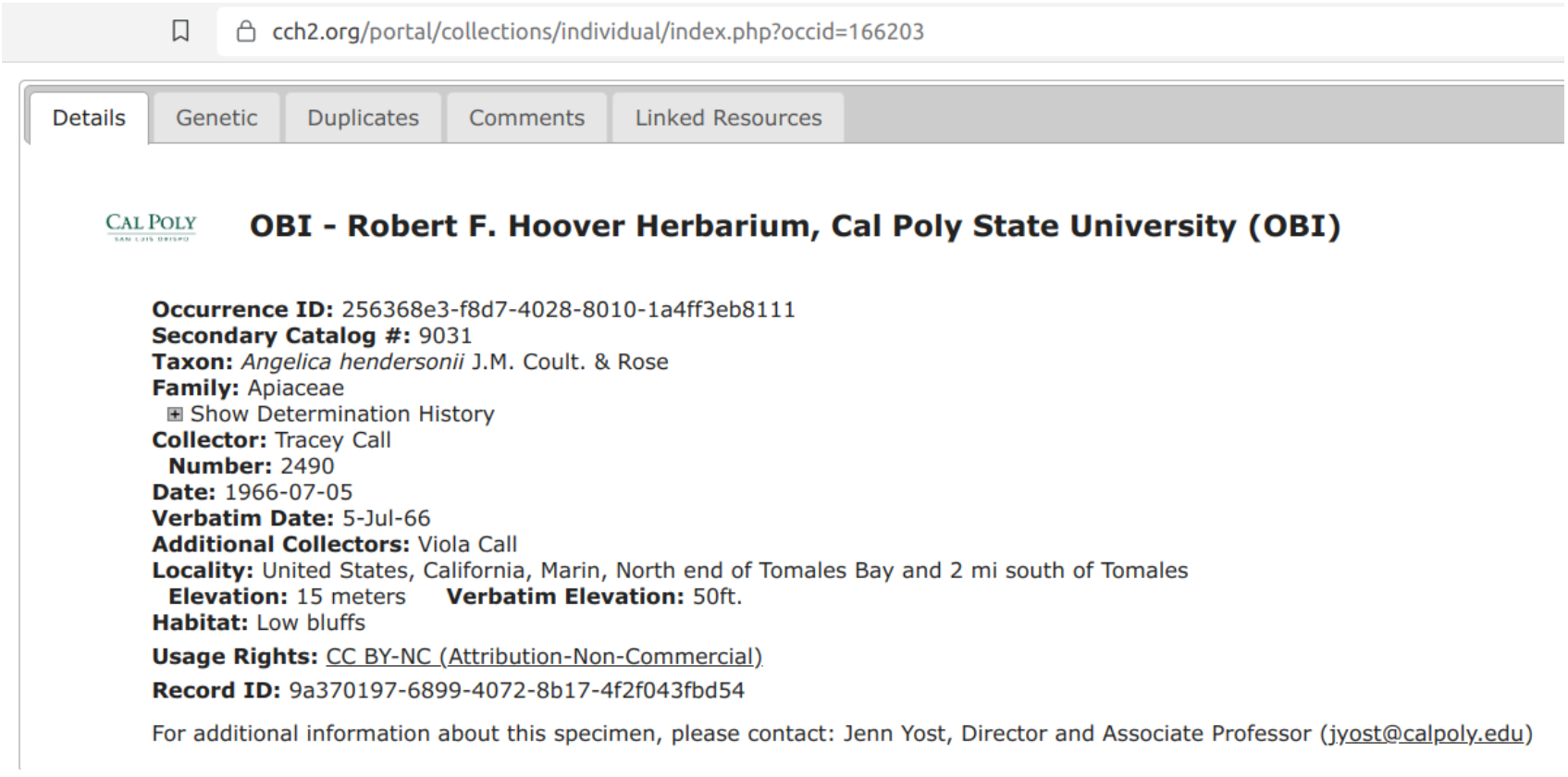


Figure 4: Herbarium Specimen **OBI09031**

Example: GenBank Accession **MT735455**

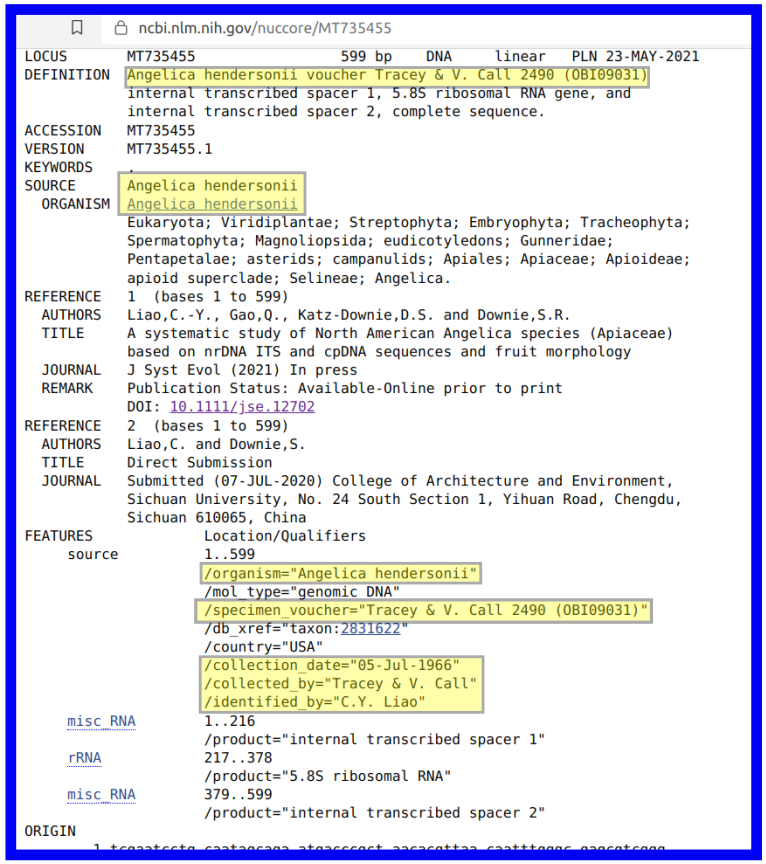
GenBank hosts a accession record **MT735455** with locus `MT735455 599 bp DNA linear PLN 23-MAY-2021` and definition *Angelica hendersonii* voucher Tracey & V. Call 2490 (OBI09031) internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence, and web reference https://www.ncbi.nlm.nih.gov/nuccore/MT735455.

Example: GenBank Accession **MsT735455**

ncbi.nlm.nih.gov/nuccore/MT735455

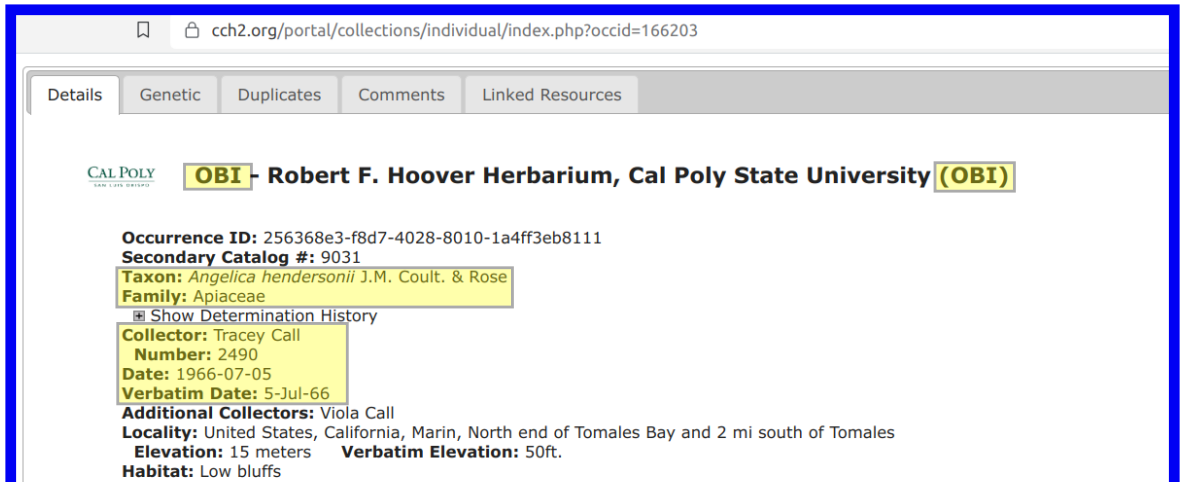
```
LOCUS       MT735455                599 bp    DNA     linear   PLN 23-MAY-2021
DEFINITION  Angelica hendersonii voucher Tracey & V. Call 2490 (OBI09031)
            internal transcribed spacer 1, 5.8S ribosomal RNA gene, and
            internal transcribed spacer 2, complete sequence.
ACCESSION   MT735455
VERSION     MT735455.1
KEYWORDS    .
SOURCE      Angelica hendersonii
  ORGANISM  Angelica hendersonii
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; asterids; campanulids; Apiales; Apiaceae; Apioideae;
            apioid superclade; Selineae; Angelica.
REFERENCE   1  (bases 1 to 599)
  AUTHORS   Liao,C.-Y., Gao,Q., Katz-Downie,D.S. and Downie,S.R.
  TITLE     A systematic study of North American Angelica species (Apiaceae)
            based on nrDNA ITS and cpDNA sequences and fruit morphology
  JOURNAL   J Syst Evol (2021) In press
  REMARK    Publication Status: Available-Online prior to print
            DOI: 10.1111/jse.12702
REFERENCE   2  (bases 1 to 599)
  AUTHORS   Liao,C. and Downie,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-JUL-2020) College of Architecture and Environment,
            Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu,
            Sichuan 610065, China
FEATURES             Location/Qualifiers
     source          1..599
                     /organism="Angelica hendersonii"
                     /mol_type="genomic DNA"
                     /specimen_voucher="Tracey & V. Call 2490 (OBI09031)"
                     /db_xref="taxon:2831622"
                     /country="USA"
                     /collection_date="05-Jul-1966"
                     /collected_by="Tracey & V. Call"
                     /identified_by="C.Y. Liao"
     misc_RNA        1..216
                     /product="internal transcribed spacer 1"
     rRNA            217..378
                     /product="5.8S ribosomal RNA"
```

Visual Inspection of OBI-GenBank Linkage



ncbi.nlm.nih.gov/nuccore/MT735455

```
LOCUS       MT735455                599 bp    DNA     linear   PLN 23-MAY-2021
DEFINITION  Angelica hendersonii voucher Tracey & V. Call 2490 (OBI09031)
            internal transcribed spacer 1, 5.8S ribosomal RNA gene, and
            internal transcribed spacer 2, complete sequence.
ACCESSION   MT735455
VERSION     MT735455.1
KEYWORDS    .
SOURCE      Angelica hendersonii
  ORGANISM  Angelica hendersonii
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; asterids; campanulids; Apiales; Apiaceae; Apioideae;
            apioid superclade; Selineae; Angelica.
REFERENCE   1  (bases 1 to 599)
  AUTHORS   Liao,C.-Y., Gao,Q., Katz-Downie,D.S. and Downie,S.R.
  TITLE     A systematic study of North American Angelica species (Apiaceae)
            based on nrDNA ITS and cpDNA sequences and fruit morphology
  JOURNAL   J Syst Evol (2021) In press
  REMARK    Publication Status: Available-Online prior to print
            DOI: 10.1111/jse.12702
REFERENCE   2  (bases 1 to 599)
  AUTHORS   Liao,C. and Downie,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-JUL-2020) College of Architecture and Environment,
            Sichuan University, No. 24 South Section 1, Yihuan Road, Chengdu,
            Sichuan 610065, China
FEATURES             Location/Qualifiers
     source          1..599
                     /organism="Angelica hendersonii"
                     /mol_type="genomic DNA"
                     /specimen_voucher="Tracey & V. Call 2490 (OBI09031)"
                     /db_xref="taxon:2831622"
                     /country="USA"
                     /collection_date="05-Jul-1966"
                     /collected_by="Tracey & V. Call"
                     /identified_by="C.Y. Liao"
     misc_RNA        1..216
                     /product="internal transcribed spacer 1"
     rRNA            217..378
                     /product="5.8S ribosomal RNA"
     misc_RNA        379..599
                     /product="internal transcribed spacer 2"
ORIGIN
```



cch2.org/portal/collections/individual/index.php?occid=166203

Details | Genetic | Duplicates | Comments | Linked Resources

CAL POLY

OBI - Robert F. Hoover Herbarium, Cal Poly State University (OBI)

Occurrence ID: 256368e3-f8d7-4028-8010-1a4ff3eb8111
Secondary Catalog #: 9031
Taxon: *Angelica hendersonii* J.M. Coult. & Rose
Family: Apiaceae
Show Determination History
Collector: Tracey Call
Number: 2490
Date: 1966-07-05
Verbatim Date: 5-Jul-66
Additional Collectors: Viola Call
Locality: United States, California, Marin, North end of Tomales Bay and 2 mi south of Tomales
Elevation: 15 meters **Verbatim Elevation:** 50ft.
Habitat: Low bluffs

Outcomes

1. Associated 25 OBI Specimen With 100 GenBank Records
2. Identified Next Collection of Interest: California Botanic Garden Herbarium (RSA)

jhpoelen.nl/obi-genbank/

Extending OBI Herbarium Records to include associated NCBI GenBank sequences

hash://md5/40b93e072ceb31bb9e78078b929f19d8
hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6

Jorrit Poelen

Katelin Pearson

Jenn Yost

2023-07-19

Abstract

Specimen from Natural History Collections are physical repositories of genetic information. Genetic sequences extracted from specimen are stored in genetic sequence databases like the openly accessible GenBank at NCBI, DNA DataBank of Japan, or the European Nucleotide Archive (ENA). While researchers and collection managers make efforts to associate (or link) Natural History Collection records with their derived genetic accession records, extra work is need to make these associations explicit. We describe how a collaboration between a biodiversity informatics expert and collection managers of the Hoover/OBI Herbarium at CalPoly, San Luis Obispo, CA was forged with the aim to extend OBI specimen records to include their associated GenBank records. In addition, we quantify the costs of creating these specimen extensions, and discuss the socio-economic capacity needed to repeat this digital specimen extension process for the hundreds of millions of specimen records available globally today.

3. Enabled *Streaming* of GenBank Records via ASU's BioKIC [a]
4. Built Prototype Workflow to Version, Filter, and Merge GenBank/DwC-A [b]

---

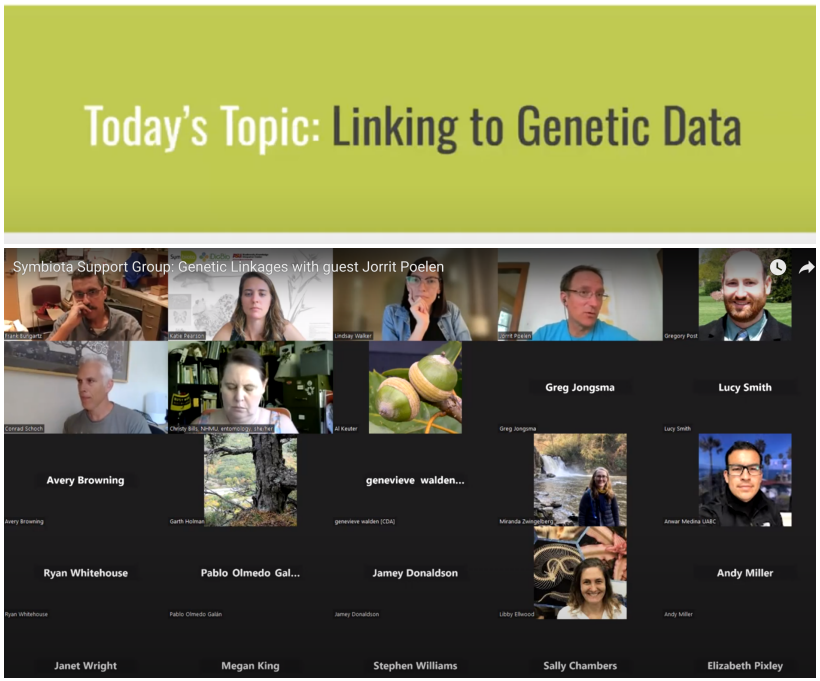
[a] Biodiversity Knowledge Integration Center @ biokic.asu.edu

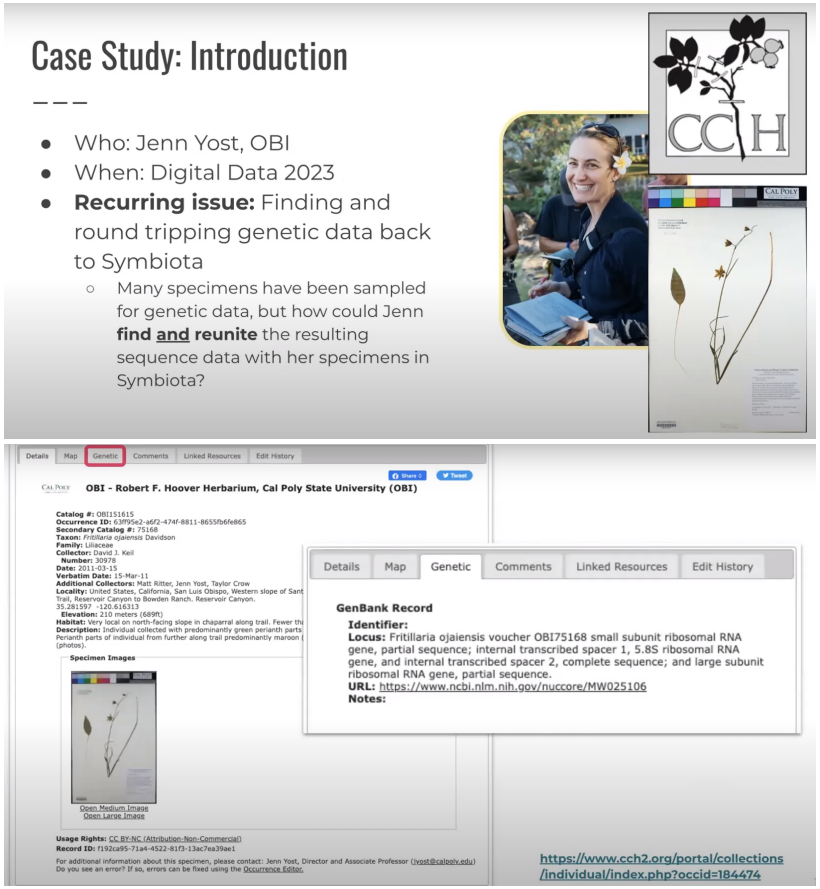
[b] Poelen et al. 2023 jhpoelen.nl/obi-genbank

Rewards and Paying It Forward

Share and Discuss

Lindsay Walker and Katie Pearson hosted a recorded SSG session organized around our experiences on linking genetic sequences [6].





[6] Symbiota Support Hub. (2023, September 11). Symbiota Support Group: Genetic Linkages with guest Jorrit Poelen. YouTube. https://youtu.be/H76eeKxECEs

Acknowledgement

Big thanks to Jenn Yost, Katie Pearson, Lindsay Walker, Nico Franz, Greg Post, and many others collaborators/supporters for their willingness to experiment and try things, new and old.

Guiding Questions

What would *Jorrit* like *others* to do to realize the Digital Extended Specimen?

1. Communicate, Collaborate, and Explore [7]
2. Remix, Reuse "Boring" Open Tools with Existing Open Data [8]
3. (Only If Absolutely Needed) Try Something New [9]
4. Goto 1. [10]

---

[7] Consider hiring specialists like Jorrit to sustain *your* work, and promote group cognition.

[8] Save the planet: recycle biodiversity data/tools/knowledge.

[9] Articulate *specific* needs, and find suitable collaborators.

[10] Keep experimenting!

Guiding Questions

What do *you* do to realize the Digital Extended Specimen?

What would you like *others* to do to realize the Digital Extended Specimen?