

# 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 <sup>1</sup>?

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

---

<sup>1</sup>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

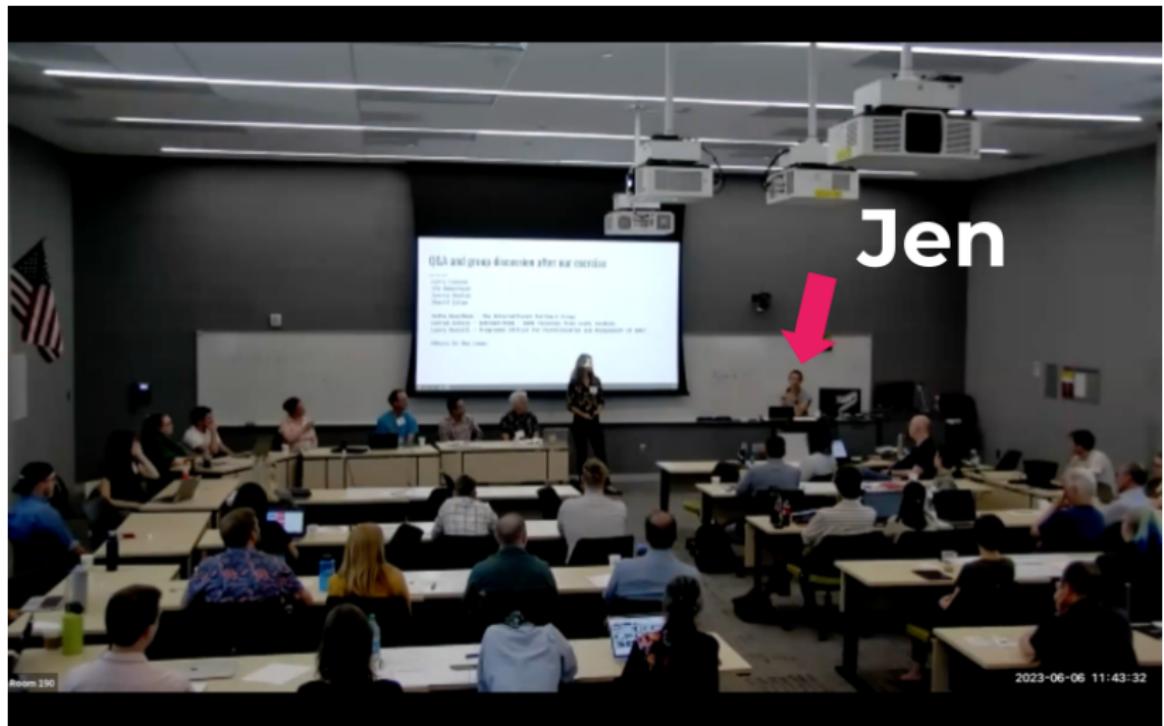
Jenn Yost, Katie Pearson, Lindsay Walker, Jorrit Poelen

**DIGITAL DATA CONFERENCE**  
LEVERAGING DIGITAL DATA FOR CONSERVATION, ECOLOGY,  
SYSTEMATICS, AND NOVEL BIODIVERSITY RESEARCH  
**HYBRID EVENT**  
**JUNE 5-7, 2023**  
EVENT LOCATION: ARIZONA STATE UNIVERSITY  
& VIRTUALLY THROUGH ZOOM  
VISIT [bit.ly/3GGj6Xw](https://bit.ly/3GGj6Xw) FOR MORE INFORMATION.

The poster has a red background with a stylized yellow and green bird illustration on the right side. At the bottom left, there are three logos: ASU Arizona State University, iDigBio (with a globe icon), and NSC (with a DNA helix icon).



# Communicate, Collaborate, and Explore

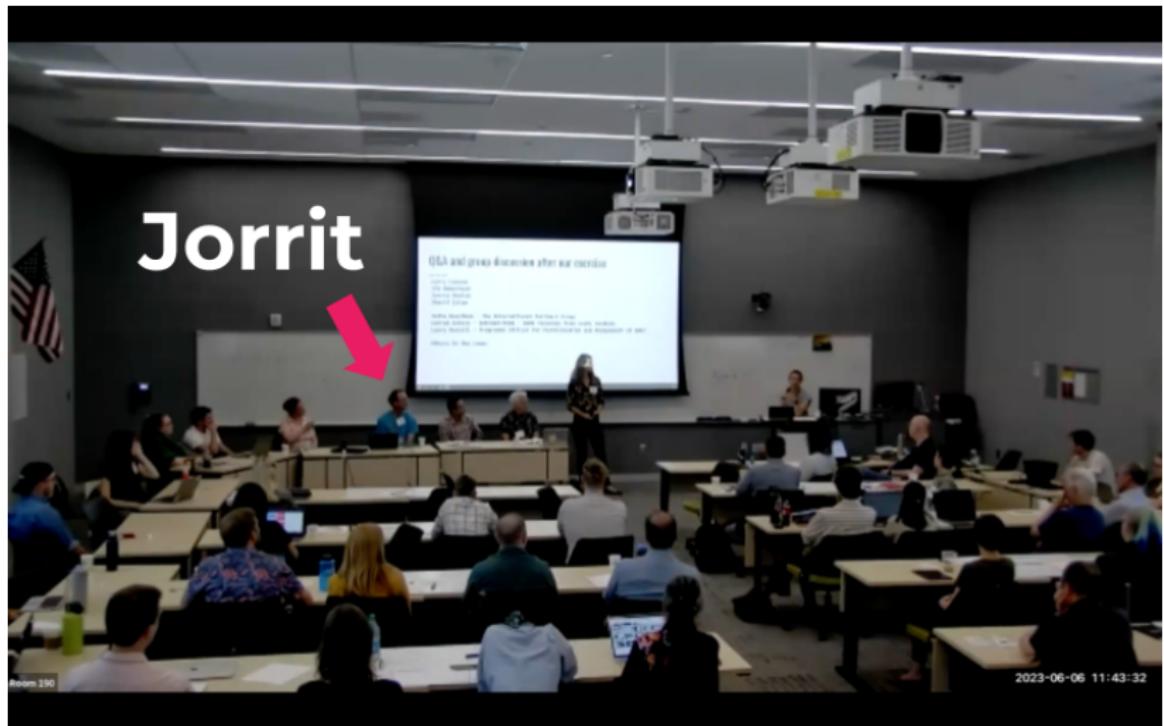


## Communicate, Collaborate, and Explore

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

— Jenn Yost 2023. [youtu.be/CNRAJvyDHu8?t=9713](https://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](https://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](https://youtu.be/CNRAJvyDHu8?t=10080)

## Method

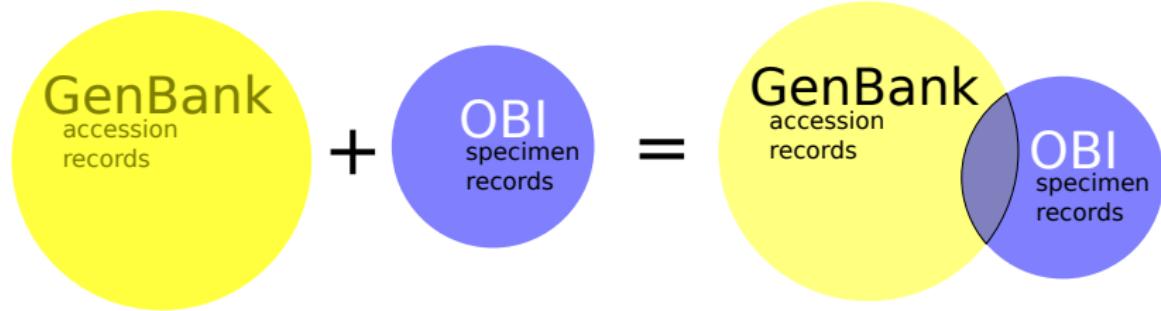


Figure 1: OBI specimen overlaps GenBank <sup>2</sup> records.

<sup>2</sup>Sayers et al. 2019. Nucleic Acids Research DOI:10.1093/nar/gky989

# Method

 National Library of Medicine  
National Center for Biotechnology Information

Nucleotide      Nucleotide ▾ OBI  
Create alert Advanced

Species      Summary ▾ 20 per page ▾ Sort by Default order ▾      Send to: ▾

Animals (216,925)  
Plants (285)  
Fungi (13)  
Protists (31)  
Bacteria (675)  
Viruses (744)  
Customize ...

Molecule types      Items: 1 to 20 of 219056

genomic DNA/RNA (147,787)  
mRNA (66,126)  
rRNA (94)  
Customize ...

Source databases      << First < Prev Page 1 of 10953 Next > Last >>

INSDC (GenBank) (39,534)  
RefSeq (179,522)  
Customize ...

Sequence Type       [Coccomyxa sp. Obi gene for 18S ribosomal RNA, partial sequence](#)  
 [Coccomyxa sp. Obi PHO-2 gene for starch phosphorylase-2, complete cds](#)  
 [Coccomyxa sp. Obi PHO-1 gene for starch phosphorylase-1, complete cds](#)

[1,899 bp linear DNA](#)  
Accession: LC473498.1 GI: 2178192011  
[Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

[5,907 bp linear DNA](#)  
Accession: LC473497.1 GI: 2178192009  
[Protein](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

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.

# Method

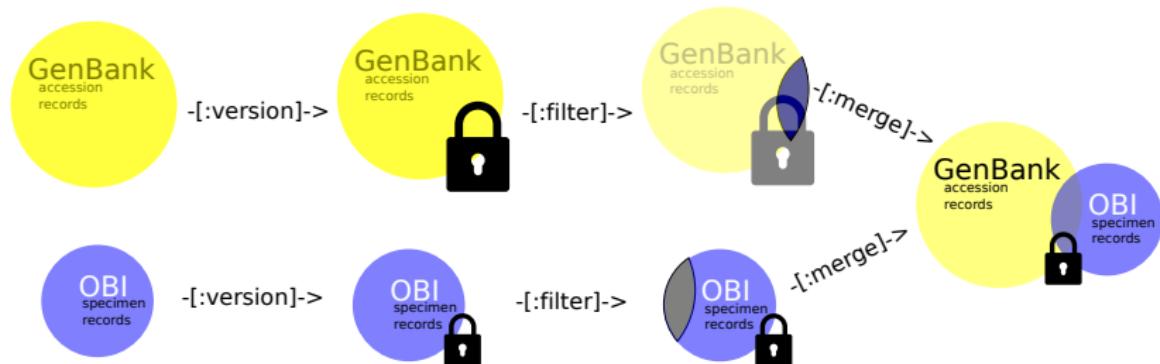


Figure 3: A workflow extending OBI specimen to include associated GenBank records enabled through versioning and data streaming using existing tools and resources (e.g., GenBank, Symbiota, Preston, grep, jq, Zenodo, BioKIC@ASU)

## Results

The Hoover Herbarium (OBI) hosts a preserved specimen of type *Angelica hendersonii* Coulter & 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>.

# Results

 [cch2.org/portal/collections/individual/index.php?occid=166203](http://cch2.org/portal/collections/individual/index.php?occid=166203)

Details    Genetic    Duplicates    Comments    Linked Resources

### 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. Coulter & 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  
**Usage Rights:** CC BY-NC (Attribution-Non-Commercial)  
**Record ID:** 9a370197-6899-4072-8b17-4f2f043fdb54

For additional information about this specimen, please contact: Jenn Yost, Director and Associate Professor ([jyost@calpoly.edu](mailto:jyost@calpoly.edu))

Figure 4: OBI09031

## Results

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>.

# Results

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	<a href="#">Angelica hendersonii</a>	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: <a href="https://doi.org/10.1111/jse.12702">10.1111/jse.12702</a>					
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 <i>/organism="Angelica hendersonii"</i> <i>/mol_type="genomic DNA"</i> <i>/specimen_voucher="Tracey &amp; V. Call 2490 (OBI09031)"</i> <i>/db_xref="taxon:2831622"</i> <i>/country="USA"</i> <i>/collection_date="05-Jul-1966"</i> <i>/collected_by="Tracey &amp; V. Call"</i> <i>/identified_by="C.Y. Liao"</i>					
misc_RNA	1..216 <i>/product="internal transcribed spacer 1"</i>					
rRNA	217..378 <i>/product="5.8S ribosomal RNA"</i>					
misc_RNA	379..500 <i>/product="internal transcribed spacer 2"</i>					

## Results

We found, among others, OBI associated GenBank accession **MT735455**, using my Lenovo T480s Laptop, BioKIC, and a 500Gb internet connection by <sup>3</sup>:

1. versioning GenBank PLN <sup>4</sup>
2. streaming all GenBank PLN records
3. including only GenBank record containing “OBI”
4. reviewing the ~200 resulting records
5. verify linked OBI specimen records

---

<sup>3</sup>Poelen JH, Pearson KD, Yost J. 2023. Extending OBI Herbarium Records to include associated NCBI GenBank sequences. <https://jhpoelen.nl/obi-genbank>  
hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cf...

<sup>4</sup>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>

# Results

□ ncbi.nlm.nih.gov/muccore/MT735455

LOCUS MT735455 599 bp DNA Linear PLN 23-MAY-2021

DEFINITION *Angelica hendersonii* voucher Tracey & V. Call 2498 (OB109031) 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; Viridiplanteae; Streptophyta; Embryophyta; Tracheophyta; Spermatophytina; Magnoliopsida; eudicots; Rosids; Asteridae; Pentapetalae; asterids; campanulids; Apiales; Apiaceae; Apioideae; apioid superclade; Selinaceae; Angelica.

REFERENCE 1. (bases 1 to 599)

AUTHORS Liao,C.-Y., Guo,O., 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.12792

REFERENCE 2. (bases 1 to 599)

AUTHORS Li,J., Liao,C.-Y., Guo,O.

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 2498 (OB109031)"  
/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 1. Angelica hendersonii (L.) M. Bieb. subsp. hendersonii (L.) M. Bieb.

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

Details Genetic Duplicates Comments Linked Resources

CAL POLY STATE UNIVERSITY LIBRARIES 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. Coul. & 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  
Habitat: Low bluffs  
Verbatim Elevation: 50ft

## Results



3. Enable *Streaming* of GenBank Records via ASU's BioKIC <sup>a</sup>
  4. Prototype Workflow to Version, Filter, and Merge GenBank/DwC-A <sup>b</sup>

<sup>a</sup>Biodiversity Knowledge  
Integration Center @ biokic.asu.edu

<sup>b</sup>Poelen et al. 2023  
[jhpoelen.nl/obi-genbank](http://jhpoelen.nl/obi-genbank)

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

## Extending OBI Herbarium Records to include associated NCBI GenBank sequences

hash://md5/40b93e072ceb31bb9e78078bb929f19d8  
hash://sha256/bc5605e58d2644baedeb16060408049f02ce528064b7fb13a5b556dd155cf6

b6

Jorrit Poelen

Katelin Pearson

Jenn Yost

2023-07-19

## Abstract

Specimens from Natural History Collections are physical repositories of genetic information. Genetic sequences extracted from specimens are stored in generic sequence databases like the openly accessible Genbank at NCBI, DNA Database of Japan, or the European Nucleotide Archive (ENA). These databases are used by researchers around the world to assess (or link) their own collections with those from other genetic studies. In addition, extra work is needed to make these associations explicit. We describe how a collaboration between a biodiversity informatics expert and collection managers of the Hoover OM Herbarium at CalPoly, San Luis Obispo, CA was forged with the aim to extend OBIS 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 specimens recorded publicly globally.

# Results



# Results

Lindsay Walker and Katie Pearson hosted a recorded SSG session organized around our experiences on linking genetic sequences<sup>5</sup>.

The screenshot shows a YouTube video player with a green header bar. The title 'Today's Topic: Linking to Genetic Data' is displayed in white text. Below the title is a grid of 15 small video thumbnails arranged in three rows. Each thumbnail shows a different person or specimen. The names of the individuals are listed below their respective thumbnails: Avery Browning, Jenn Yost, Greg Jongma, Lucy Smith, Ryan Whitehouse, Pablo Olmedo Gal., Jamey Donaldson, Andy Miller, Janet Wright, Megan King, Stephen Williams, Sally Chambers, and Elizabeth Paley.

The screenshot shows a 'Case Study: Introduction' page. It features a large image of a woman smiling while holding a plant specimen. To the right is the logo for 'CCH' (California Plant Herbaria). Below the image is a list of bullet points:

- Who: Jenn Yost, OBI
- When: Digital Data 2023
- Recurring issue: Finding and round tripping genetic data back to Symbiota
  - Many specimens have been sampled for genetic data, but how could Jenn **find and reunite** the resulting sequence data with her specimens in Symbiota?

Below this is a screenshot of a computer screen displaying a detailed view of a 'Case Study' record in a database. The record is for 'OBI - Robert F. Hoover Herbarium, Cal Poly State University (MBO)'. The 'Details' tab is selected, showing information such as 'Collection #': MBO-121508, 'Determinant': *Fritillaria affinis* (L.) Benth. ssp. *affinis*, 'Taxon': Fritillaria, 'Common Name': Yellow Fritillary, 'Family': Liliaceae, 'Order': Liliales, 'Genus': Fritillaria, 'Species': *affinis*, 'Author': (L.) Benth., 'Year': 1835, 'Voucher Date': 21 Mar 2014, 'Abstract': 'Specimen collected from a single plant growing in a shaded area in a dense forest. The plant has yellow flowers and a bulbous root system.', 'Notes': 'This specimen was collected by Robert F. Hoover, a herbarium volunteer at the Robert F. Hoover Herbarium, Cal Poly State University, San Luis Obispo, CA, USA. The specimen was collected from a single plant growing in a shaded area in a dense forest. The plant has yellow flowers and a bulbous root system.', 'Photographs': 'None', 'Image': 'None', 'Image URL': 'https://www.cch2.org/portal/collections/Individuals/index.php?objid=184476'}

<sup>5</sup>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 <sup>6</sup>
2. Remix, Reuse “Boring” Open Tools with Existing Open Data <sup>7</sup>
3. (Only If Absolutely Needed) Try Something New <sup>8</sup>
4. Goto 1. <sup>9</sup>

---

<sup>6</sup>Consider hiring specialists like Jorrit to sustain *your* work, and promote group cognition.

<sup>7</sup>Save the planet: recycle biodiversity data/tools/knowledge.

<sup>8</sup>Articulate *specific* needs, and find suitable collaborators.

<sup>9</sup>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?