

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?

¹Hardisty 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

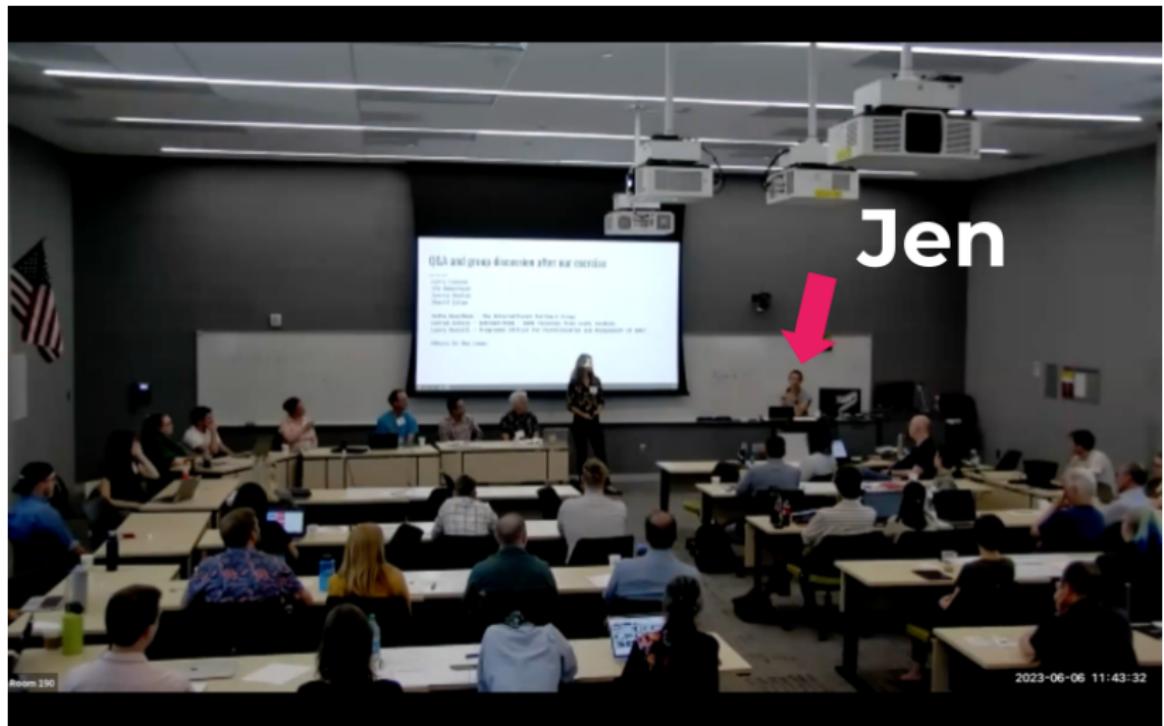
Jenn Yost, Katie Pearson, Lindsay Walker, Jorrit Poelen

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 FOR MORE INFORMATION.

The poster has a red background with a stylized yellow and green bird illustration on the right. Logos for ASU (Arizona State University), iDigBio (Integrated Digitized Bi�lioteca), and NSC (Natural Sciences Collections Alliance) are displayed at the bottom left.



Communicate, Collaborate, and Explore

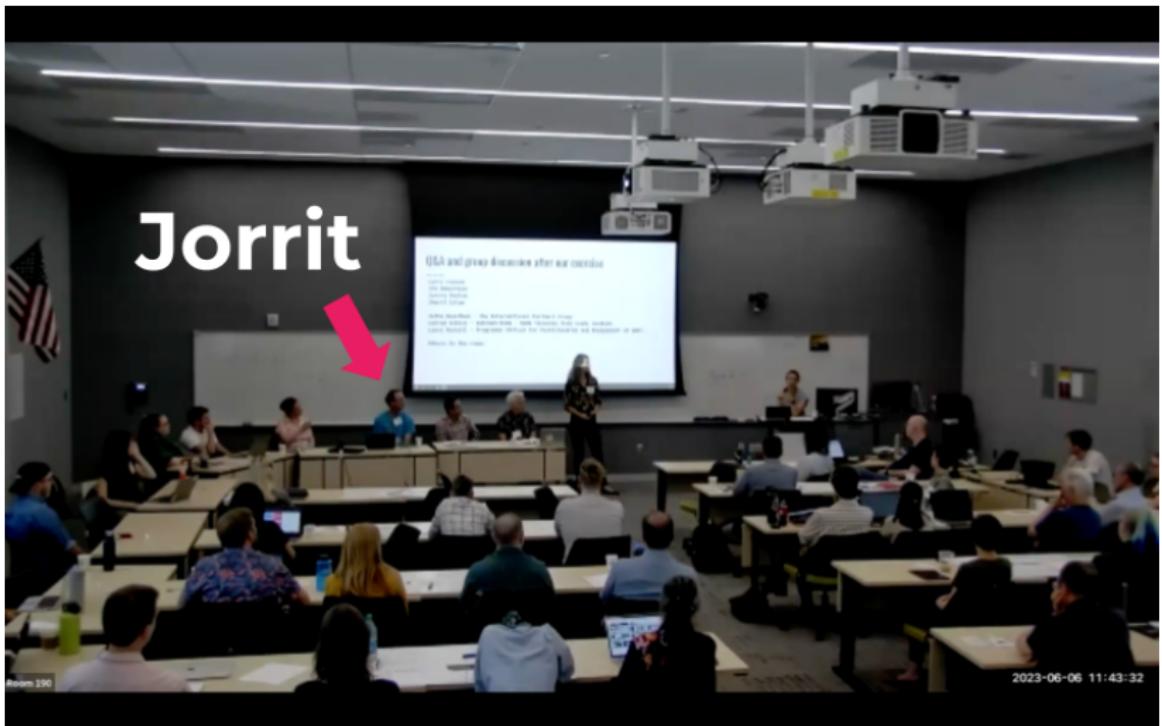


Communicate, Collaborate, and Explore

[...] 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

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”

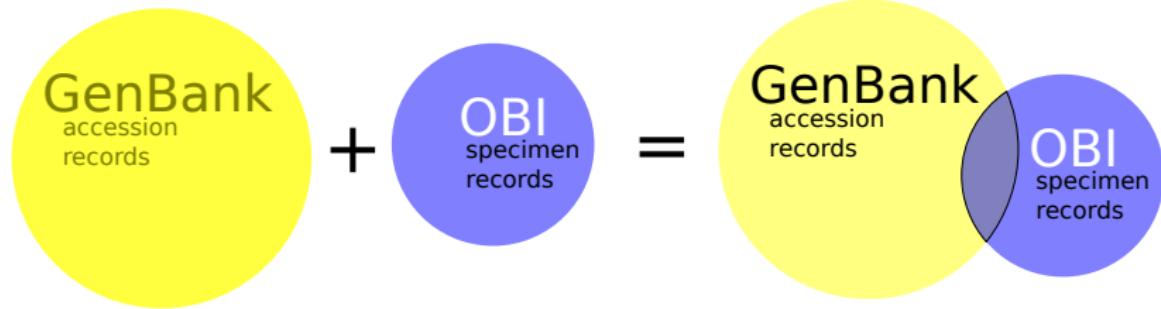


Figure 1: OBI specimens overlap GenBank ² records.

²Sayers 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

Dec., 1946 :

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.*

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

Trying Existing Tools

 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](#)

[Nucleotide](#) / [240 DEGs](#)

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

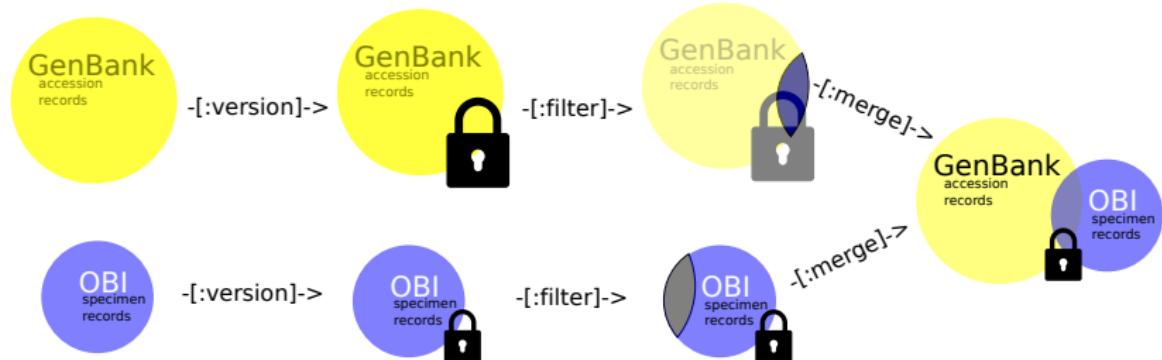


Figure 3: A workflow extending OBI specimen to include associated GenBank records enabled through versioning and data streaming using existing tools and resources (e.g., GenBank, Symbiota, Preston, grep, jq, Zenodo, BioKIC@ASU)

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

⁴Poelen JH, Pearson KD, Yost J. 2023. Extending OBI Herbarium Records to include associated NCBI GenBank sequences. <https://jhpoelen.nl/obi-genbank>
hash://sha256/be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cf...

⁵Poelen, 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* 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>.

Example: Herbarium Specimen OBI09031

The screenshot shows a web page for a herbarium specimen. At the top, there is a header with a logo and the URL cch2.org/portal/collections/individual/index.php?occid=166203. Below the header is a navigation bar with tabs: Details (which is selected), Genetic, Duplicates, Comments, and Linked Resources. The main content area displays the following information:

CAL POLY
San Luis Obispo

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: 9a370197-6899-4072-8b17-4f2f043fb54

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

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; apiod clade; 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	source	Location/Qualifiers
		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..500

Visual Inspection of OBI-GenBank Linkage

□ ncbi.nlm.nih.gov/muccore/MT735455

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

DEFINITION *Angelica hendersonii* voucher Tracey & V. Call 2498 (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; Viridiplanteae; Streptophytina; Embryophytina; Tracheophytina; Spermatophytina; Eudicots; eudicotyledons; Commelinidae; 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 Liao,C.-Y., Guo,O., Katz-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 2498 (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"
217..378
/product="5.8S ribosomal RNA"
379..599
/product="internal transcribed spacer 2"

RNA

misc_RNA

ORIGIN 1. Internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2; complete sequence.

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

Details Genetic Duplicates Comments Linked Resources

CAL POLY 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

Outcomes

The screenshot shows two main parts of the OBIS (Open Biological Observatory) interface. On the left, a detailed specimen record for a plant specimen from the Cal Poly State University (OBIS) collection is displayed. It includes fields for specimen number (OBIS-0000000000), date (2010-01-01), location (California, USA), and a photograph of the herbarium specimen. On the right, a detailed view of the associated GenBank record for the same specimen is shown. This record includes the GenBank identifier (MHW025106), a detailed description of the sequence (large subunit ribosomal RNA gene, partial sequence; Internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence), and a URL to the NCBI page (<https://www.ncbi.nlm.nih.gov/nuccore/MHW025106>). A note at the bottom of the GenBank record states: "Note: This record is associated with a specimen record in OBIS. If you have further along this record, please contact the author or the specimen's owner." Below the GenBank record, there is a "GenBank Images" section showing a small thumbnail of the herbarium specimen.

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

^aBiodiversity Knowledge Integration Center © biokic.asu.edu

^bPoelen et al. 2023
jhpoelen.nl/obi-genbank

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://md5/4093a072cc031bbfc7a0f78fb929ff19d8
hash://sha256/be5605e58d2644bbaecd1606940680d9f02ce528064b7bb13a5b556dd55cfe

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 Open Biological Observatory (OBI), DNA Data Bank of Japan (DDBJ), or the European Nucleotide Archive (ENA). While most collections utilize management systems (such as iNaturalist or Leksi), Natural History Collection records with their derived genetic accession records, extra work is needed to make these associations explicit. We describe how a collaboration between a biodiversity informatics expert and collection managers of the California OBI Herbarium at CalPoly, San Luis Obispo, worked with the National Center for Biotechnology Information (NCBI) 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.

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 sequences⁶.

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 thumbnail images. The thumbnails include various people, plants, and specimens, likely related to the topic of genetic linkages.

The screenshot shows a 'Case Study: Introduction' page. It includes a photo of a woman smiling, a logo for 'CCB H', and a detailed description of a specimen record for 'OB1 - Robert F. Hoover Herbarium, Cal Poly State University (OB1)'. The record details a specimen of *Fritillaria* with specific information about its morphology and genetic data. A 'Details' tab is selected, showing a 'GenBank Record' with a URL to the record on the CCBH portal.

⁶Symbiota 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?