

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

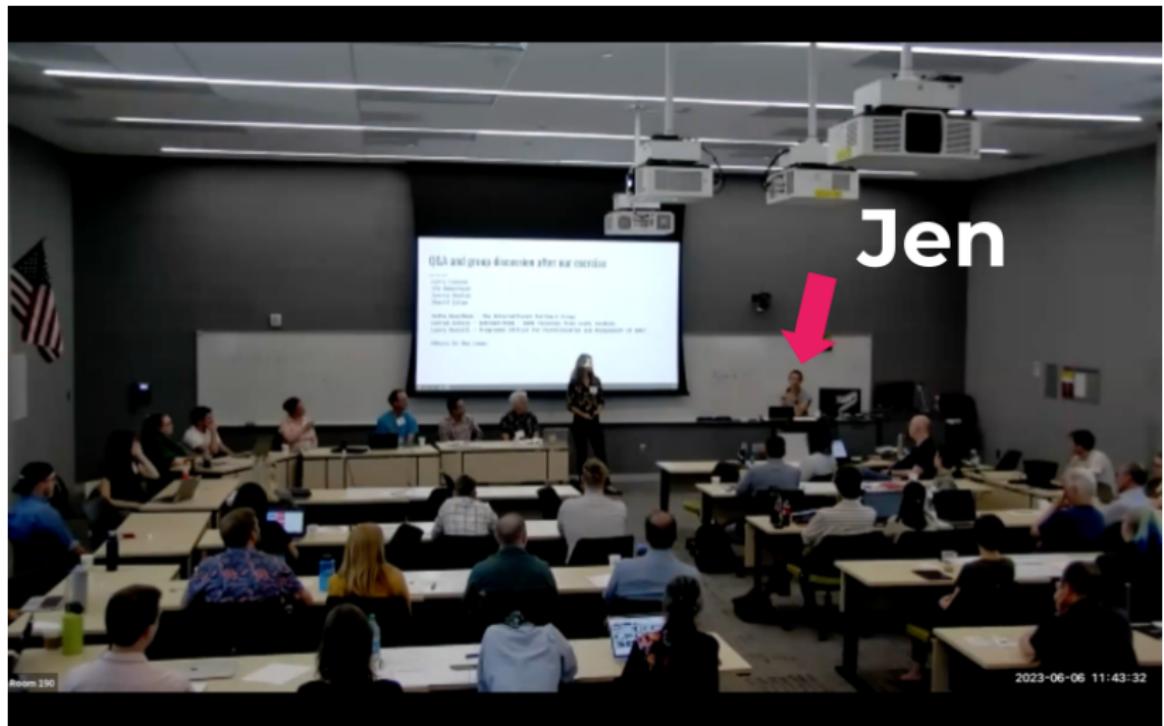
Jenn Yost, Katie Pearson, Lindsay Walker, Jorrit Poelen

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 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

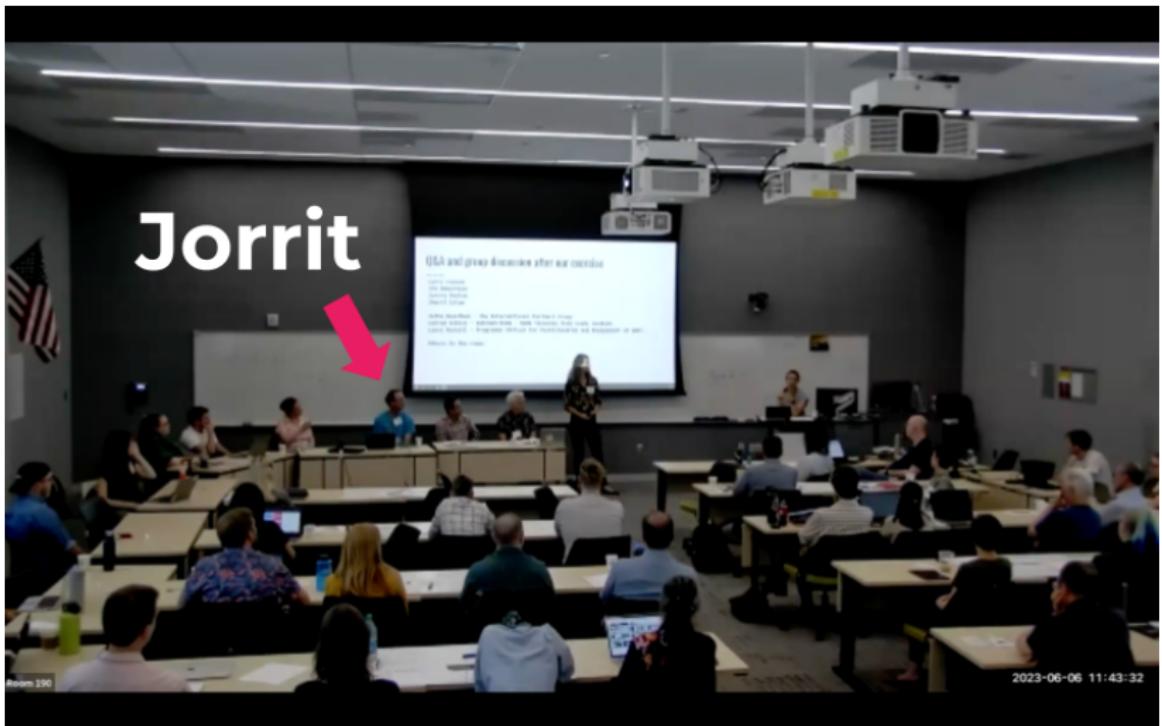


Communicate, Collaborate, and Explore

[...] 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

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

Methods

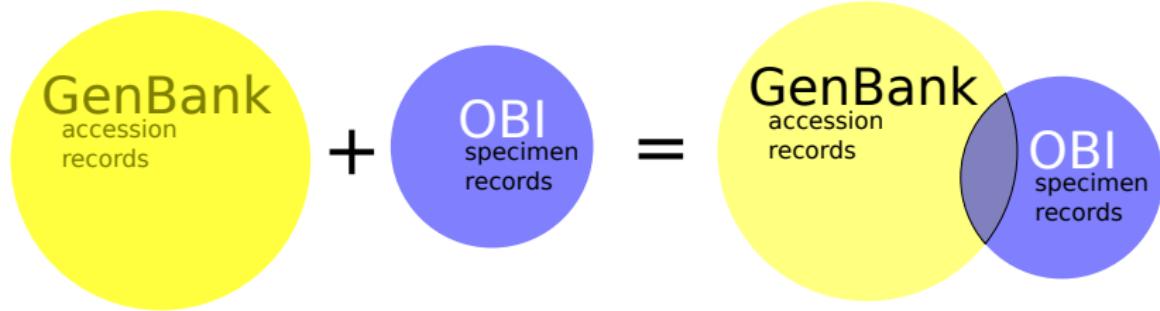


Figure 1: OBI specimen overlaps GenBank ² records.

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

Methods

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

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

Sequence Type Nucleotide / 240,063

Items: 1 to 20 of 219056 << First < Prev Page 1 of 10953 Next > Last >>

[Coccomyxa sp. Obi gene for 18S ribosomal RNA, partial sequence](#)
1. 1,899 bp linear DNA
Accession: LC473498.1 GI: 2178192011
[Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Coccomyxa sp. Obi PHO-2 gene for starch phosphorylase-2, complete cds](#)
2. 5,907 bp linear DNA
Accession: LC473497.1 GI: 2178192009
[Protein](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)

[Coccomyxa sp. Obi PHO-1 gene for starch phosphorylase-1, complete cds](#)

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.

Methods

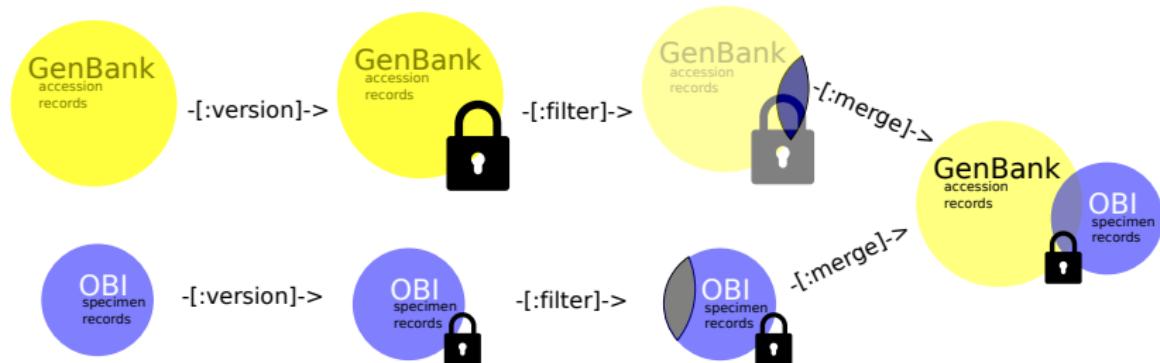


Figure 3: A workflow extending OBI specimen to include associated GenBank records enabled through versioning and data streaming using existing tools and resources (e.g., GenBank, Symbiota, Preston, grep, jq, Zenodo, BioKIC@ASU)

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

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)

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	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 <i>/organism="Angelica hendersonii"</i> <i>/mol_type="genomic DNA"</i> <i>/specimen_voucher="Tracey & 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 & 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 ³:

1. versioning GenBank PLN ⁴
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

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

⁴Poelen, 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/nucleotide/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; Angelicaceae.

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. Internal transcribed spacer 1
2. 5.8S ribosomal RNA
3. Internal transcribed spacer 2

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

Details Genetic Duplicates Comments Linked Resources

CAL POLY STATE UNIVERSITY LIBRARIES 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

The screenshot shows two main panels. The left panel displays a specimen record for a plant specimen from the Cal Poly State University (OBIS) collection. It includes details like specimen number, date, collector, and location. The right panel shows the corresponding 'GenBank Record' for the same specimen, listing the identifier, accession number, and sequence details. A thumbnail image of the herbarium specimen is also visible.

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

- Enabled Streaming of GenBank Records via ASU's BioKIC ^a
- 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

jhpoelen.nl/obi-genbank/

Extending OBI Herbarium Records to include associated NCBI GenBank sequences

hash://md5:4093c072cc031bbfc78078b029f19d8
hash://sha256:be5605e5bd264baecd1606940680d9f02ce528064b7bb13a5b556dd55fcfe

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 Access GenBank (OAG), DDBJ Database of Japan, or the European Nucleotide Archive (ENA). While most collections utilize management systems (such as iNaturalist), 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 OBIS 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.

Results



Results

Lindsay Walker and Katie Pearson hosted a recorded SSG session organized around our experiences on linking genetic sequences⁵.

Case Study: Introduction

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?

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

Catalog #: DBI75168
Collection date: 2010-05-20
Secondary Catalog #: 75168-00000000000000000000000000000000
Type Specimen: No
Paratype: No
Material Type: Plant
Date: 2010-05-20
Author: Robert F. Hoover
Locality: Loma Linda, California, San Bernardino County, San Bernardino Co., CA, USA
Altitude: 1000 m
Abstract: This specimen is a chapter story leaf. Power 04
Description: Individual specimen with undetermined gross morphological characters. Leaflet with serrated margin, petiole with long pubescence.

Specimen Images:

Geotag Record:

Identifier: Locus: Fritillaria esculentum voucher DBI75168 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, partial sequence; large subunit ribosomal RNA gene, partial sequence

URL: <https://www.ncbi.nlm.nih.gov/sviewer/viewer.fcgi?acc=NM022106>

Notes:

Usage Rights: CC-BY-NC-ND, Attribution-NonCommercial

Review ID: 9155557 (Last Review: 2023-05-22)

The information contained in this record is provided by the author(s) and Associate Professor, ccbh@calpoly.edu.
Do not use or copy it if you do not have the permission of the copyright holder.

<https://www.ccbh.org/portal/collections/Individuals/index.php?ocid=1844676>

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

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

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

⁸Articulate *specific* needs, and find suitable collaborators.

⁹Keep 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?