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 Jenn Yost Katelin Pearson

2023-09-11

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.

Hello again Lindsay, Jenn, Katie,

Thinking about our Extended Specimen Workshop / OBI-GenBank collaboration kept my poor little brain quite active today and yesterday. In fact, I had trouble sleeping because of it.

As I am stewing on an abstract for your 2023-09-11 Symbiota Support Hub session, I reflected on the effort it took for us to nurture our collaboration:

1. time investment to organize extended specimen workshop at Digital Data 2023 at Tempe, AZ
2. time investment to package NCBI GenBank and OBI Herbarium digital resources (Poelen, Pearson, and Yost 2023; Poelen 2023)
3. time investment to build a custom (off-line enabled) workflow based on (archived) digital resources of known origin

4. time investment to archive the NCBI GenBank Plant flat files at ASU's BioKIC via Globus facilitated by Greg Post and Nico Franz
5. time investment by Katie and Jorrit to collaborate on a shared google sheets to propose (Jorrit) and verify (Katie) GenBank<>OBI association claims
6. time investment by Katie to populate OBI's Symbiota records with associated GenBank sequences
7. transfer of symbolic reward (a tub of TJs Ginger Cookies) by Symbiota developer Ed Gilbert at the July 2023 workshop on imagining a Biological Action Center. This workshop itself required a 3 day time investment on my part.

and now . . .

(pending)

8. More time investment (mostly by Jorrit) to publicize a novel workflow to discover GenBank associations in existing natural history collections as published through DwC-A.

In an effort to do a little cost/benefit analysis, I made a quick back-of-the-napkin calculation of the (socio-)economics aspects of our experiment: I spent about 16 hours of work (I kept track of my time, this excludes writing this text) and got a tub of 78 TJs Ginger Cookies (13 servings at 6 cookies a serving). (Thank you!) Noting that the resale value of TJ Ginger Cookies is probably about $1 or less, I favor using cookies as a unit instead of a dollar. So, converting that to an hourly effort would: 78 cookies / 16h ~ 5 cookies / hour. Another way quantify the "value" of our method is to estimate the number of cookies gained per created specimen GenBank association (as measured from https://cch2.org/portal/content/dwca/OBI_DwC-A.zip with signature hash://sha256/cd9de973510975dac3394952bba9c486a482762b3beab05ecb678037b99ab85b as seen on 2023-07-19T14:46:11.145Z):

78 cookies / 25 GenBank associations = 3 cookies / OBI-GenBank association

Assuming that someone is willing to work for 3 cookies per association, and assuming that OBI is representative collection as far as 0.03% of specimen (i.e., 94,031 OBI preserved specimen) having GenBank associations (25 detected genbank associations), and estimating about 200 million digitized preserved specimen (GBIF claims 225M, iDigBio claims 138M as of 2023-08-10), you'd have to buy = 200M * 0.03% * 3 = 60k associations * 3 cookies / association = 180k cookies or about 2300 tubs of TJs Ginger Cookies.

The case for making our method to efficiently produce/distribute the number of cookies needed per GenBank association:

1. monitor the availability of specimen-GenBank links in DwC-A and GenBank (link out) over time

2. estimate time needed to discover new, and maintain existing, specimen-GenBank links
3. produce re-usable methods, reducing development time
4. advertise the method to avoid rework
5. improve methods when needed through available means (standarization, "smart" algorithms, efficient semi-automated link suggestion/curation workflows)
6. express the value of having specimen-GenBank links readily/openly available

So, now my open questions are:

Q1. How can we estimate the methods needed to support the discovery and maintenance of Specimen-GenBank records such that it can be sustained by those valueing the availability of Specimen-GenBank records links?

This is taking into account that one person cannot possibly eat 180k cookies before they go stale, even if someone is willing (and able) to source the 2300 cookie tubs needed to discover the specimen-genbank claims. So, distribution of the cookies should be factored into the development of the Specimen<>GenBank association discovery and recording method.

As it stands, the work that Jorrit has done to showcase a method to extend digital representations preserved specimen with their GenBank associations is not yet valued in terms of "real" monetary units. Which means that his work is, economically speaking, valueless.

Q2. What is the value of Jorrit's work so far? And who should compensate him? What is 16 hours of my time worth to you? A $10 tub of cookies?

And, perhaps more importantly,

Q3. do the current funding mechanisms allow for rapid development of ideas to address immediate needs in the biodiversity informatics community?

I'd like very much to contribute to your support hub event, and before doing so, I'd like to discuss the thoughts above. The reason for taking on this prototype challenge was to gather some evidence to support the idea that current funding mechanisms and project management are insufficient to build something as dynamic and complex as the digital extended specimen.

It make me think of a proverb I first encountered as a 6 year old, and found inspirational only later in life:

> "Everybody want to go back to nature, but nobody wants to walk."

which can be reworded in terms of the digital extended specimen:

> "Everybody want to have the digital extended specimen, but someone else has to build it."

This may be a bit extreme of a statement in context of our collaboration, especially because I know that you've invested plenty of time in extending existing digital specimen with their GenBank sequences and beyond.

In short, I'd like to have more discussion around effective collaboration that help nurture, and sustain, the socio-economical aspects of the digital extended specimen. Without these productive collaborations, the massive amount of work needed to better integrate the biodiversity / biology and informatics disciplines continues to be as is - sporadic and mostly based on volunteer work. A similar, but more broad argument, can be made for creating good working conditions to promote quality research (Rahal et al. 2023).

Curious to hear your thoughts,

thx, -jorrit

References

Poelen, Jorrit H. 2023. "GenBank PLN (Plantae, Fungi, Algae) Sequence Index in TSV, CSV, JSONL Formats Hash://Sha256/Bc7368469e50020ce8ae27b9d6a9a869e0b9a2a0a9b5480c69ce6751f Hash://Md5/F6f78f64e3b3ff06adc3229badbd578b." Zenodo. https://doi.org/10.5281/zenodo.8117720.

Poelen, Jorrit H., Katelin Pearson, and Jenn Yost. 2023. "Extending OBI Herbarium Records to Include Associated NCBI GenBank Sequences. Hash://Sha256/Be5605e58d2644baedcb160604080d9f02ce528064b7fbb13a5b556dd55cfeb6." GitHub. https://github.com/jhpoelen/obi-genbank.

Rahal, Rima-Maria, Susann Fiedler, Adeyemi Adetula, Ronnie P.-A. Berntsson, Ulrich Dirnagl, Gordon B. Feld, Christian J. Fiebach, et al. 2023. "Quality Research Needs Good Working Conditions." *Nature Human Behaviour* 7 (2): 164–67. https://doi.org/10.1038/s41562-022-01508-2.