Speaker Notes to

Module 4: Sharing: making library, archive & museum collections FAIR

"Case study: collections-as-[FAIR]-data in a marine science library"

Amanda Whitmire, Stanford University Libraries

Slide 2:

Before I get into a few examples of the kinds of things I'm working on, I want to offer some personal / professional context. I used to be an oceanographer, and now I'm a librarian at a marine lab. I embrace the collections-as-data approach to collection development b/c is it critical to being able to provide my community (faculty, staff, students, external researchers) with materials they would use. They need information in formats that are actionable. The currency of observational data in the natural sciences is tabular data. Or at least digital text. We put so much work into scanning and accessioning, creating metadata for our collections, but I suspect that PDFs will get very little use.

One of the core principles is, "Collections as data … stewards should … work to lower the barriers to use…" This is an evolution in how librarians approach our work that I personally think is critical. For the community I support, digitization is not enough. It doesn't lower the barrier to use. I must engage in transformation to be doing work that is truly useful to researchers. In the context of marine science, there is lack of ecologically and climatologically relevant observational data. To be able to assess long-term changes, like shifts in biodiversity due to changing ocean conditions, researchers need data over long time periods. Hopkins Marine Station has been here for over 100 years. Through the digitization and transformation of our physical collections, the data in our library can help fill observational gaps for our researchers, so that's what we're trying to do. What I want to convey is the sense that we, as cultural heritage professionals, should be interrogating how we are processing collections and how we're offering them. Sometimes scanning is enough. Sometimes it's not, and when it's not, let's all be curious about what tools we can use or develop to get the transformations we need.

Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data https://zenodo.org/records/3066209
Vancouver Statement on Collections as Data (2023) https://zenodo.org/doi/10.5281/zenodo.8341519

Slide 3:

I only have a few minutes to offer you what I think are good examples of how we could demonstrate how we apply FAIR principles to collections, which I personally view as meaning that we make our collections actionable into research products. I'm going to give two VERY QUICK, high-level examples, but I have way more and could go into a lot more detail - this is just a sample.

Hopkins Marine Station was founded in 1892 and was the first marine station on the west coast of the US. Our offshore oceanographic sampling began in 1928 with Henry Bryant Bigelow, continued from 1929 – 1937 with "Station C" formally established over the Monterey Submarine Canyon. Sampling

started up again in 1950 when Hopkins got itself a research vessel. These datasets form the beginning of the oceanographic record for Monterey Bay.

See more: https://exhibits.stanford.edu/station/feature/hydrobiological-survey-of-monterey-bay

Slide 4:

So, as an example here, I'm showing a data report that we had on our shelves that had literally almost 100 pages of data tables that included ocean temperature, salinity, oxygen, phosphate, and silicate. It shares weekly sampling at a single station in Monterey Bay, from the surface down to 800 meters, 1951-1955. We scanned the report, transcribed the data, first with OCR using FineReader and then double-checking every row of data by hand, and shared the dataset in Stanford Digital Repository as a CSV file. Is the data FAIR? Eh, more or less. What would make it more FAIR? Depositing the data in the World Ocean Database, where it will be available in context, completely interoperable with other oceanographic measurements.

Slide 5:

Other datasets available as a part of this same sampling program, which ran from 1951 – 1978, include phytoplankton and zooplankton diversity and abundance, along with the oceanographic data. Coming soon: Weekly oceanographic data at 6 stations in the Bay, running 1954 – 1978.

Slide 6:

And, there's WAY more oceanographic data and biodiversity data in other collections we have that we are in various stages of making FAIR. The Stanford Oceanographic Expeditions ran from 1963-1973 and they were pretty much constantly on the water collecting data in one form or another – there is a LOT of work to do with these materials. We have scanned everything and added metadata for discovery, but haven't yet transformed any of it into public datasets.

Slide 7:

Before I wrap up, I'm going to toss out one more example that is maybe more accessible to other kinds of librarians – turning thesis observations into FAIR data.

This story starts with a dissertation, and continues to today, and many of you will be familiar with this work. In 1931, a graduate student at Hopkins named Willis Hewattbegan a survey of invertebrates along a transect line at Hopkins Marine Station. The transect was 105 yards long, and Hewattidentified and counted the 90 most common species on the transect from October 1931- June 1933. His dissertation is in the stacks at Miller Library.

Sarah Gilman and Rafe Sagarin were undergraduate students at Hopkins in 1993. For their spring class research project, they chose to resample 19 of the Hewatttransect quadrats and compare their observations with Hewatt's findings. What they found was that the species abundances had shifted toward more southern species, that is, to warmer water species. Hopkins also has a daily temperature record going back to 1919, and they speculated that the species shift that they observed between

1933 and 1993 was the result of warming ocean temperatures. They wrote up this work in a final paper, which I have in the stacks at Miller Library.

Rafe and Sarah returned to gather more samples and eventually published a paper in Science about the effect of climate change on intertidal species diversity. Rafe carried on sampling the transect and it was a central part of his dissertation work, and then he still kept coming back to monitor the transect after he his career trajectory took him to Duke and then the University of Arizona. Hopkins researchers continue the time-series, which is now referred to as the Hewatt-Sagarin transect.

Slide 8:

I began to wonder, what would it look like to treat our collection of Hewattmaterials as data? Well, step number 1 was to scan the dissertation, which we did. https://searchworks.stanford.edu/view/2151287. Remember that I said that one of the central principles of Collections as Data is that stewards aim to lower barriers to use. It's well understood that our researchers work digitally, so after we scanned the dissertation, we looked for what we could extract and offer as standalone data that would be of interest to marine ecologists.

For the Hewattdissertation, my digital collections specialist, Melissa, extracted 19 data tables, including a list of species observed along the transect provided by Hewatt, which is data that people are very interested in. What is so important about old data like this is how it offers an important baseline, or at least historical context, for marine intertidal biodiversity. There is absolutely no way to assess whether ecosystems are changing *now* without having information about what they were like *before*. So, the big deal about Hewatt's work is his observations and data about what kinds of things were seen in the intertidal while he was sampling out there in the 1930s, and we turned it into a dataset that's now in SDR.

Slide 9:

After all of the data tables and species lists were transcribed, I compiled and deduplicated a list of every taxonomic name that Hewattobserved at Hopkins Marine Station, and set about verifying the accuracy of the names, some of which were misspelled or have changed over time and connected them with taxonomic authorities. I was able to use a fantastic open-source tool called Global Names Finder, which can detect even misspelled taxonomic names (within reason). Global Names is now available as a reconciliation endpoint in OpenRefine, which is a free web-based tool that you can use to clean data. Once I verified the names, I was able to connect with GBIF, WikiData, and WoRMSto enhance the dataset with identifiers for these commonly used and authoritative databases. GBIF—the Global Biodiversity Information Facility — holds over 2 billion occurrence records and is the worlds largest biodiversity database. Wikidataand Wikipedia pages have descriptive information and link out to photographs, references, and other taxonomic identifiers that are part of the biodiversity knowledge graph. WoRMS, the World Register of Marine Species, has much more taxonomic information about each species, including vernacular names, synonyms, basionyms, the full classification, and so on. Then we can export this from OpenRefineto include with our dataset as another potentially useful derivative from the Hewattdissertation.

Slide 10:

To summarize, instead of scanning and posting the thesis in the repository being the whole of the story, it was really just the first step toward treating the thesis in a “collections as data” way, of trying to make the the thesis into data, and then making the data FAIR. It’s a lot of work, but I believe it’s the kind of work we should be doing.