Sabrina She

LIS 545 | Mayernick

2/27/26

Term Project - Draft of Final Submission

Contents:

Updated Data and Metadata Profile - Emilio Levier Natural History Specimens

The data I am focusing on for this profile is a set of data about the natural history specimens collected by Emilio Levier in the mid to late 1800s. All specimens were collected from France, Georgia, Italy, Portugal, Spain, or Switzerland. Though collected by Levier, the dataset was originally created in Bionomia by Joaquim Santos. Key stakeholders in this data include Bionomia, a database of natural history specimen collections, which the dataset is drawn from (as well as on Global Biodiversity Information Facility AKA GBIF). This set is also indexed on OpenAIRE. Though Levier himself has long passed, Joaquim Santos is also a stakeholder here as the creator. Other notable stakeholders include researchers, natural history and specimen organizations, and the various institutions that the

specimens were originally recorded with. This dataset includes two files. Within those files, there are the scientific names of the specimens, country and location of origin, the institute which the specimens were originally recorded with, and various identification keys and other information. The files come in both CSV and JSON formats, and my device, which is a Macbook, automatically opens them in either Numbers or TextEdit, which are both fairly universal programs on a computer. Additionally, the data itself is listed as open and publicly accessible, however, it is also restricted within Creative Commons rights. Both files contain the same data, and are simply formatted differently. The JSON file makes examining each specimen easier by compiling all the information into one chunk of text, while the CSV file is better for gaining a large overview of all the data from its spreadsheet format.

This dataset includes the following metadata: name/title, subject, creator, a short description, the dates of creation, keywords and subject tags, format, rights, publisher name, source, and an identifier in the form of a DOI. The metadata can be found most easily interpreted in the CSV file, as well as in the site page for the dataset. The same metadata is present in the JSON file, but it's slightly more difficult to read at a glance. While fairly comprehensive, the provided metadata lacks some descriptiveness, which impacts its discoverability for unfamiliar users. For example, though the files are titled, they are both simply "Q4256884.filetype", which doesn't inform users on the contents of the files.

In terms of structure, the dataset appears to use something close to Ecological Metadata Language (EML), which lists GBIF as one of its main users, and GBIF is also the host of the specimen data which this dataset pulls from. I make the assumption of EML being used based on the metadata

being pretty similar to the required elements that the GBIF lists on their guide, but they're slightly adapted to fit the time period in which Levier was recording his specimens. The actual webpage for the dataset is standardized across the Zenodo repository, with most of them having the same base elements and information. Like I mentioned above, the lack of descriptiveness brings down the browsability of the dataset. I also think that adding locations and dates to the subject tags would help to increase this, as well as the overall category of specimens collected. For the actual title of the dataset, I noticed that many other datasets included the name of the subjects studied, like "Glacier mass data" or "Empathy in Pediatric Nursing Education", which worked well to draw me in as a user, increasing my likelihood of reading and engaging with that data. Another helpful element to add would be a 'Read me' file, explaining the codes inside the files so that users who aren't familiar with this type of dataset are more easily able to understand the data on a line by line level. Lastly, though not strictly necessary, I think it would be beneficial to add a 'related datasets' feature in tandem with the existing 'citations' section at the bottom. This would be helpful in showing users what similar datasets exist, and what other researchers have used them for to hopefully give an example of what their own work could look like.

Because this dataset is so new, no publications based on this dataset are currently listed on Zenodo. However, the Bionomia page for Levier's profile contains a 'Science Enabled' tab that shows the user a list of publications which have used Levier's work as a citation or base for their research. There are currently 107 works listed under that tab. A quick search of Levier's name in GBIF pulls up six published works, which cite his research in each respective publication. Through his Wikidata page, I am easily able to find a list of organizations that hold pieces of his work in their collections, on whose own databases I would no doubt find more citations of him.

WORKS CITED

Bionomia. (n.d.-a). https://bionomia.net/

Emilio Levier. Wikidata. (n.d.). https://www.wikidata.org/wiki/Q4256884

EML (ecological metadata language). EML (Ecological Metadata Language) – Metadata Standards Catalog. (n.d.). https://rdamsc.bath.ac.uk/msc/m16

GBIF. (2011, April). *Gbif metadata profile: How-to guide version 1.0 April 2011*. gbif.org. https://www.gbif.org/sites/default/files/gbif_resource/resource-80641/gbif_metadata_profile_how-to_en_v1.pdf

GBIF. (n.d.-b). https://www.gbif.org

Levier, E. (2026, January 28). *Natural history specimens collected and/or identified and deposited.* Zenodo. https://zenodo.org/records/18406059

Mayernik, M. (2020, March 16). *Metadata (Isko Encyclopedia of KO)*. Metadata (IEKO). https://www.isko.org/cyclo/metadata#4.1

Updated Repository Profile - Smithsonian National Museum of Natural History

The repository I've chosen for this data set is the Smithsonian National Museum of Natural History (Link: https://www.re3data.org/repository/r3d100013822). Their digital collections already cover very similar subject matter, and they have a separate set of Department of Botany collections, which Levier's data would fit perfectly into. Levier's data also appears to contain information matching physical specimens, which the Smithsonian also happens to collect. This is particularly suitable because both formats of data could be housed within the same repository, making his work more accessible to researchers. While this repository is open to all, the Smithsonian holds a set of requirements geared towards research and educational purposes. Solicited acquisitions are highly encouraged, but all submissions and acquisitions remain subject to their donations and collections management policy.

In general, the repository is open for submissions. The collection scope for the Department of Botany falls within botanical specimens and related information. The museum will accept donations that have a known origin and/or are solicited. The acquisition is also likely dependent on the content matter and whether or not the museum contains data that is too similar in any of its millions of existing collections. Though not explicitly stated, the data will likely also have to include some metadata to describe the collection, such as names, dates, and origin, included in the 'known origin' requirement. The museum heavily discourages unsolicited donations, especially ones that have been shipped to them without prior discussion. They do not list limitations on data types, but for the Department of Botany, the only data types they show are specimens that are either photographed, drawn, or contain text information.

While the SIP was unable to be found, the repository does provide their criteria for acquiring specimens (under 'Acquisition' in the Department of Botany section), as well as general criteria and policies for acquisitions and collections management. It appears as though to obtain this package and others, a submitter would have to first contact the Department of Botany, who would then consider if they were interested in considering the dataset for submission. The digital Botany collection contains just over 4.5 million records, so I would imagine this is a very selective process.

A large part of that selection process is the acquisitions, and within the Smithsonian's policies, all of the acquisition methods listed require at least one point of human contact. They range from bequesting, which can be more hands-off, to the very involved field work acquisition. The first step in a donor submitting a request is to email the appropriate department for their data, so that is the very first human assistance they encounter, which then rolls into the consultation step.

According to their website, the Smithsonian uses IPTC (International Press Telecommunications Council) standards for all digitized data. It follows that the Smithsonian would like metadata to be submitted according to this structure, but may also be submitted in a form that can be easily adapted. For example, all that they ask for in their submissions section is for standard information like name, origin, type, date, and so on, citing that they prefer to display this information to the public to encourage access and proper stewardship. The full list of display information is listed on the museum's collection policies page.

When using the digital collection, there is no login process required to download the data. Options for downloading include exporting files as CSV or KML (Keyhole Markup Language) for text files. There is no direct download button for images, but it is easy to right click and select 'save image'

to download the image for botanical drawings and photographs. Each item also contains an EZID (Easy ID) linking back to the item's exact page for ease of reference. For many of the photos, there does not appear to be an easy way to download the metadata. On the other hand, if a researcher were to request a dataset, it is likely that the museum would provide the metadata in a downloadable file.

Finally, while a Smithsonian-specific DIP could not be found, this is how a DIP was defined in a document listed as a resource under the Smithsonian archival page: "The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS." (CCSDS, 2002). AIP here is Archival Information Packet, and this resource goes on to describe that archivists or those in charge of presenting the information to a researcher will assemble this package upon researcher request.

WORKS CITED

Consultative Committee for Space Data Systems. (2002, January). *Reference Model for an Open Archival Information System (OAIS)* . Smithsonian Institution Archives. https://siarchives.si.edu/sites/default/files/pdfs/650x0b1.PDF

Bradyh. (2017a, April 27). *Digital curation*. Smithsonian Institution Archives. https://siarchives.si.edu/what-we-do/digital-curation

Levier, E. (2026, January 28). Natural history specimens collected and/or identified and deposited. Zenodo. https://zenodo.org/records/18406059

Museum collections policies. Smithsonian National Museum of Natural History. (n.d.).

https://naturalhistory.si.edu/research/nmnh-collections/museum-collections-policies

Smithsonian Institution. (2023, September 18). *Collections Management Policy*. National Museum of Natural History.

https://naturalhistory.si.edu/sites/default/files/media/file/2024cmp508final.pdf

Smithsonian National Museum of Natural history. (n.d.).

https://www.re3data.org/repository/r3d100013822

Smithsonian National Museum of Natural history. Botany Collections Search. (n.d.).

https://collections.nmnh.si.edu/search/botany/

Recommended Data Citation

She, Sabrina *"Emilio-Levier: Natural history specimens collected and/or identified and deposited."* Github, Levier, Emilio & Santos, Joqauin. Zenodo, 2/27/26. https://github.com/shesa257/Emilio-Levier

Preservation Considerations

This dataset is protected under Creative Commons and is hosted through Zenodo, but is also available through Bionomia and OpenAIRE, so the multiple avenues of publication make the files less likely to suddenly become inaccessible, and the duplication in itself helps to protect against deterioration. Both file formats are fairly standard and would not cause any issues with current technology. The CSV file is particularly resilient given its basic format. However, if technology or the standard file format were to dramatically shift for whatever reason, both files would need reformatting and/or migration to different host repositories. Both files require specific software to open, but that software is fairly standard across all computers. The JSON file opens in Excel, Sheets, or Numbers, and the CSV file opens in a plain text editor. It is highly unlikely that either types of software would become obsolete so quickly that the files couldn't be modified or updated in time.

Copyright License Statement

The original dataset in Zenodo is under the Creative Commons license, which I believe is fully appropriate. If we were to get specific, I think a CC BY-NC-SA license is the most suited to a scientific dataset such as this one. It is a Creative Commons license where the creator must be credited (the BY part) that allows for only non-commercial use (that's the NC portion), and adaptations are permitted but must be non-commercial and credited as well (the SA part). This is better than a CC BY-NA or CC BY-SA because adaptations will be necessary as data is reformatted to fit each researcher's project, which could count as an adaptation, but still protected against being used for profit and the crediting requirement ensures that Emilio Levier and Joaquin Santos are given proper acknowledgement for all their hard work.

Human Subject Considerations

This dataset contains little to no personally identifiable information besides the names of the creators. The data itself is regarding botanical specimens, so no human protection is needed there. Outside of that scope, there is the potential for the data to be unethical in regards to how it was collected and where the specimens are physically being stored. Namely, I am thinking of the data and specimens being recorded without permission, or if there is a larger issue like the travelling of non-native plants that could become invasive. Of course, these problems are external to the realm of the dataset and human considerations.