(DRAFT) Fairy Tales and Digital Research Data

Jorrit H. Poelen

2025-11-17

PRO
POLL
SOIL

Presented at the 17-19 Nov 2025 kick-off meeting in Freisingen, Germany of the EU Horizon's ProPollSoil project "Understanding and managing soil health impacts to protect soil-dependent pollinators" https://doi.org/10.3030/101219108.

Adapted from:

Poelen, J. H., & Seltmann, K. C. (2025, October 23). Book Binding for the Digital Age hash://md5/097a516df8d5eb03aadcaeed942e557b. Datos Vivos/Living Data 2025, Bogotá, Colombia. Zenodo. https://doi.org/10.5281/zenodo.17392711 a presentation as part of the Datos Vivos conference in Bogotá, Colombia 21-24 Oct 2025.

Cite As

Poelen, J.H. (2025) Fairy Tales and Digital Research Data. Zenodo. https://doi.org/10.5281/zenodo.

License

Guiding Questions

How to keep our Digital Biodiversity Data around?

How to make best use of our (past, present, future) digital research data?

What is **one** of **your** experiences that comes to mind in which digital data played a *not* so beneficial role in a transnational, multi-institutional, collaborative research community? Share your horror story!

Digital Data on the Internet are Ephemeral

> *[. . .] We began in 1996 by archiving the Internet itself, a medium that was just beginning to grow in use. Like newspapers, the content published on the web was ephemeral - but unlike newspapers, no one was saving it. [. . .]* [1]

[1] Internet Archive. 2025. Accessed on 2025-10-13 at https://archive.org/about

Digital **Biodiversity** Data on the Internet are Ephemeral

[. . .] 20%-75% of biodiversity datasets in data networks GBIF, iDigBio, DataONE, and BHL changed or were unavailable in 2019/2020.[. . .] [2]

[2] Elliott et al. 2020. Ecol Inf. doi:10.1016/j.ecoinf.2020.101132

Who or what keeps our Digital Biodiversity Data around?

a) **the Data Fairy.**

b) Awareness of the fragility of digital data.

c) Common sense data archiving and citation practices.

Who or what keeps our Digital Biodiversity Data around?

a) **the Data Fairy.**
Who is this Data Fairy?

b) Awareness of the fragility of digital data.

c) Common sense data archiving and citation practices.

Is GBIF our Data Fairy?

GBIF Secretariat provides a publication **framework** for biodiversity data, but **is neither the owner nor custodian of such data**, and therefore is not responsible for the actual content served by Data Publishers.

GBIF Secretariat cannot guarantee the quality or completeness of data, **nor does it guarantee uninterrupted data access services**. Users employ these data and services at their own risk. [3].

[3] https://www.gbif.org/terms/data-user as accessed on 2025-10-13

Is GBIF our Data Fairy?

GBIF Secretariat provides a publication **framework** for biodiversity data, but **is neither the owner nor custodian of such data**, and therefore is not responsible for the actual content served by Data Publishers.

GBIF Secretariat cannot guarantee the quality or completeness of data, **nor does it guarantee uninterrupted data access services**. Users employ these data and services at their own risk. [4].

No!

GBIF is not claiming to be a data fairy who keeps your digital archives.

[4] https://www.gbif.org/terms/data-user as accessed on 2025-10-13

Is Zenodo our Data Fairy?

[. . .] Data files are versioned. Records are not versioned. [. . .] Records can be retracted from public view; however, the data files and record are preserved. [. . .] Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. [. . .] All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. [. . .] In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories. [. . .] [5].

[5] https://about.zenodo.org/policies/ as accessed on 2025-11-05

Is Zenodo our Data Fairy?

> [. . .] Data files are versioned. Records are not versioned. [. . .] Records can be retracted from public view; however, the data files and record are preserved. [. . .] Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. [. . .] All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. [. . .] In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories. [. . .] [6].

No!

Zenodo is not claiming to be a data fairy who keeps your digital archives.

[6] https://about.zenodo.org/policies/ as accessed on 2025-11-05

Data Fairy is a Cousin of the Poop Fairy.

Data Fairy is a Cousin of the Poop Fairy.

example.org/data

404 Not Found

Recent Reminders of Data Fairy Absence

- Server infrastructure running the Symbiota Hosted Portals at Arizona State University (ASU) was compromised and taken offline 21 Jul 2025 and was said to be fully restored on 10 Oct 2025 after hard work by the Symbiota Support Hub team. The outage affected over 54 hosted collections.
- The Symbiota Collections of Arthropods Network (**SCAN: https://scan-bugs.org**) serving specimen occurrence records and images from over 100 North American arthropod collections for all arthropod taxa, was taken offline in 2025 with no plans to revive it.
- 20 years since its inception and facilitating access to over 300k digitized biodiversity data works, the **Biodiversity Heritage Library** faces an uncertain future as it is set to lose their institutional sponsor, the Smithsonian, on 1 Jan 2026 [7].

[7] https://about.biodiversitylibrary.org/about/future-of-bhl/

Who or what keeps our Digital Biodiversity Data around?

a) ~~the Data Fairy~~

b) **Awareness of the fragility of digital data.**

c) Common sense data archiving and citation practices.

Who or what keeps our Digital Biodiversity Data around?

a) ~~the Data Fairy~~

b) Awareness of the fragility of digital data.

c) **Common sense data archiving and citation practices.**

Book Binding for Digital Biodiversity Data

- What is so neat about physical books?
- What is book binding for biodiversity data?
- Example 1: The iDigBio Archives
- Example 2: Plazi's BHL Corpus
- Example 3: GloBI's Archive and Review of SCAN

The Neat Thing about Physical Books

Even before the invention of the book press, books and scrolls have been pretty successful in transferring knowledge across generations and around the world.

Typically, books are portable stacks of bound paper containing text and imagery.

Books can combined into collections without changing their design.

Books are kept around the world in (little) public libraries, academic institutions, private collections and national archives.

Books are wireless, their content cannot be easily altered remotely, changes can be detected (ripped out pages), and they need no power to operate.

Books can be sent by physical mail.

The Neat Thing about Physical Books

Even before the invention of the book press, books and scrolls have been pretty successful in transferring knowledge across generations and around the world.

Typically, books are portable stacks of bound paper containing text and imagery.

Books can combined into collections without changing their design.

Books are kept around the world in (little) public libraries, academic institutions, private collections and national archives.

Books are wireless, their content cannot be easily altered remotely, changes can be detected (ripped out pages), and they need no power to operate.

Books can be sent by physical mail.

Idea . . . what if we treat digital data more like a bound book instead of a web location?

Steps to Binding Digital Data Books

Step 1. Use **signed** citations to reference digital data [8].

Step 2. Bind these referenced data by describing them in a **Data Bill of Materials** (DataBoM).

Step 3. Publish the data bill of materials, a digital text file.

Step 4. Use the signed citation of the data bill of material (DataBoM) in your research.

Step 5. Continuously monitor the availability of the DataBoM and the associated data.

[8] Elliott et al. 2023. Sci Data. doi:10.1038/s41597-023-02230-y

Example 1: DataBoM for iDigBio Data Registry

Create an iDigBio Data Bill of Materials by capturing their registered datasets, and describing their origins,

```
(iDigBio Registry with Institutional DwC-A Data URLs)
  -[:take snapshot and download DwC-As]
     ->(DataBoM + DwC-A files)
```

using the following Preston [9] command

```
preston track --seed https://idigbio.org
```

[9] https://github.com/bio-guoda/preston

Data Bill of Material (DataBoM) in English

Expressing the digital content and their origin of the DataBoM in “plain” English:

“A version of the iDigBio registry was downloaded on 2025-10-01 from <...idigbio.org/v2/search...> with content signature <hash://sha256/52d6...>. This iDigBio registry version had member dataset urn:uuid:650... associated with <.../UCSB-IZC_DwC-A.zip> . And this DwC-A URL had content signature <hash://sha256/3d4e...> as seen on 2025-10-01.”

Data Bill of Material (DataBoM) in rdf/nquads

or, made more machine readable using Provenance Ontology [10] and Hash URIs [11] as expressed in rdf/nquads:

```
<...idigbio.org/v2/search...>
  <hasVersion>
    <hash://sha256/52d6...> .
<hash://sha256/52d6...>
  <hadMember>
    <urn:uuid:650...> .
<urn:uuid:650...>
  <hadMember>
    <.../UCSB-IZC_DwC-A.zip> .
<.../UCSB-IZC_DwC-A.zip>
  <hasVersion>
    <hash://sha256/3d4e...> .
```

[10] https://www.w3.org/TR/prov-o/

[11] Elliot et al. 2023. Sci Data. doi:10.1038/s41597-023-02230-y

DataBoM Binds Data Together

As a text file, the DataBoM has a content signature that uniquely identifies the digital bound collection of signed data it references.

So, retrieval method for data bundle defined by DataBoM with signature X is:

1. get the DataBoM by their signature X.
2. get data listed in DataBoM by their signatures.

Note that we are asking for the data content, not the data location. Also, content signatures are format agnostic, so any content (of any size) can be included.

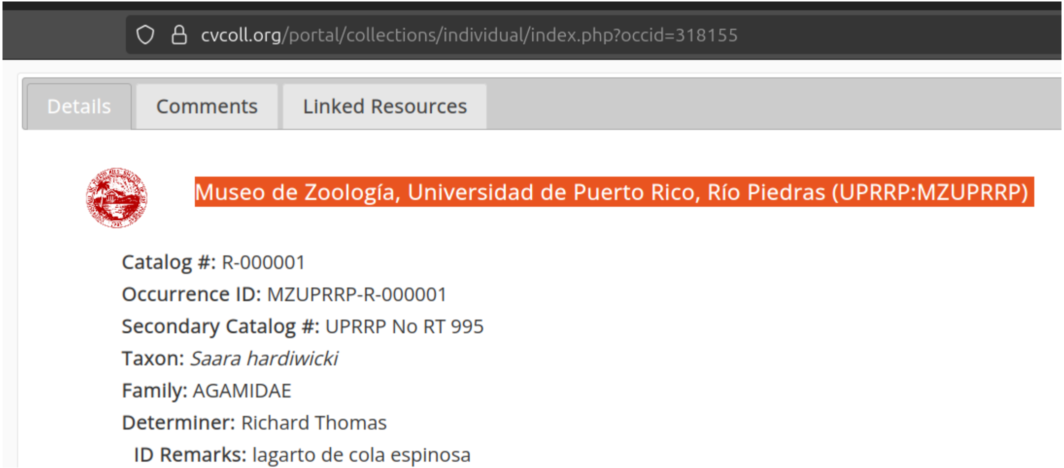
DataBoM Binds Data Together

Here's a retrieval method for the first DwC record in the data bundle defined by DataBoM with a sha256 signature starting with 40c4... as **expressed in a bash script**.

```
preston cat\
 --remote https://linker.bio \
 hash://sha256/40c44d75d243e\
8d1fde2376483637df6f96bfe182\
bb4bcd119cb5311cfdbc000 \
 | preston dwc-stream \
  --remote https://linker.bio \
 | head -1
```

producing the first record as *Saara hardiwicki*, a lizard specimen from Pakistan from Museo de Zoología, Universidad de Puerto Rico, Río Piedras (UPRRP:MZUPRRP).

We found a lizard specimen from Pakistan in a Puerto Rican Collection!



cvcoll.org/portal/collections/individual/index.php?occid=318155

Museo de Zoología, Universidad de Puerto Rico, Río Piedras (UPRRP:MZUPRRP)

Catalog #: R-000001

Occurrence ID: MZUPRRP-R-000001

Secondary Catalog #: UPRRP No RT 995

Taxon: *Saara hardiwicki*

Family: AGAMIDAE

Determiner: Richard Thomas

ID Remarks: lagarto de cola espinosa

DataBoM Flexibility

Data Bill of Materials (DataBoMs)

1. ... can reference other DataBoMs, including older versions of a DataBoM.
2. ... can be tiny or super huge
3. ... allow for arbitrarily detailed description of data provenance in natural language or using structured text like rdf/nquads in combation with the provenance ontology.

With this, DataBoMs are a recursive data structure allowed to grow to arbitrary size.

Examples 2. and 3. — More Existing DataBoMs

>300k Digitized Biodiversity Heritage Literature Items

Poelen, J. H., & Agosti, D. (2025). A Versioned Literature Corpus derived from Biodiversity Heritage Library hash://md5/b3cd9de0685deeebf57a5d225e59c10f (0.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.16616872

Darwin Core Archives Associated with scan-bugs.org

Elton, Nomer, & Preston. (2025). Versioned Archive and Review of Biotic Interactions and Taxon Names Found within globalbioticinteractions/scan hash://md5/58de50154e330c331993fe5d0852ad84. Zenodo. https://doi.org/10.5281/zenodo.16894884

Who or what keeps our Digital Biodiversity Data around?

a) ~~the Data Fairy~~

b) **Awareness of the fragility of digital data.**

c) **Common sense data archiving and citation practices.**

How to keep our Digital Biodiversity Data around?

I've been using this Data Bill of Materials approach to track terrabytes of biodiversity data since Sept 2018, keeping identical copies across independent storage locations, and reusing their content.

How do *you* intend keep our digital heritage around?

And Remember

example.org/data

404 Not Found

Thank you!

Made possible (in part) by Horizon Europe's 101219108, NSF's DBI 2027654, DBI 2102006 and OAC 1839201.

For questions/comments/ideas, please do reach out to:

Jorrit H. Poelen

https://jhpoelen.nl

jhpoelen@jhpoelen.nl

https://orcid.org/0000-0003-3138-4118

Guiding Questions

How to keep our Digital Biodiversity Data around?

How to make best use of our (past, present, future) digital research data?

What is **one** of **your** experiences that comes to mind in which digital data played a *not* so beneficial role in a transnational, multi-institutional, collaborative research community? Share your horror story!