

# Book Binding for the Digital Age

Format Agnostic Method to Review and Archive Biodiversity Data

Jorrit H. Poelen    Katja C. Seltsmann

2025-10-23

Presented as part of the Datos Vivos conference in Bogotá, Colombia 21-24 Oct 2025.



DATOS  
VIVOS  
2025

## Cite As

Poelen, J.H.; Seltmann, K.C. (2025) Book Binding for the Digital Age. Zenodo.  
<https://doi.org/10.5281/zenodo.17352156>

## License

CC BY 4.0. For license text, see <https://creativecommons.org/licenses/by/4.0/>.



# 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. [...]*<sup>1</sup>

---

<sup>1</sup>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.[...] <sup>2</sup>*

---

<sup>2</sup>Elliott et al. 2020. Ecol Inf. doi:10.1016/j.ecoinf.2020.101132

# Why does our Digital Biodiversity Data Keep Disappearing?

Theory 1. People believe in the Data Fairy.

# Why does our Digital Biodiversity Data Keep Disappearing?

Theory 1. People believe in the Data Fairy.

*Who is the Data Fairy?*

Data Fairy is a Cousin of the Poop Fairy.





Data Fairy is a Cousin of the Poop Fairy.



## Recent Reminders of Data Fairy Absence

- ▶ **iDigBio** supported server infrastructure running the Symbiota Hosted Portals at ASU were 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 NSF grant supporting iDigBio ends in 2026 <sup>3</sup>.
- ▶ 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 <sup>4</sup>.

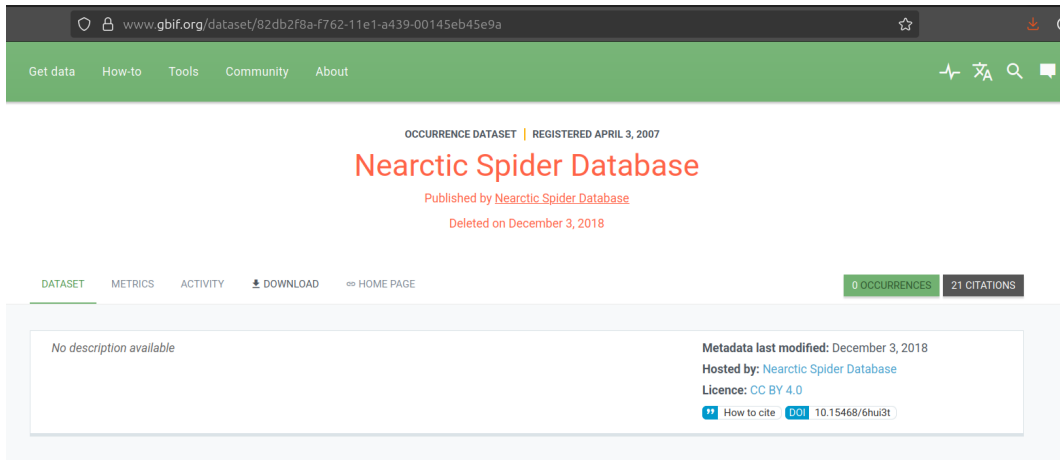
---

<sup>3</sup>[https://www.nsf.gov/awardsearch/showAward?AWD\\_ID=2027654](https://www.nsf.gov/awardsearch/showAward?AWD_ID=2027654)

<sup>4</sup><https://about.biodiversitylibrary.org/about/future-of-bhl/>

# Another Notable Reminder of Data Fairy Absence

The Nearctic Spider Database is lost after a flooded basement in 2010. Their web domain was abandoned and is now serving ads for medical marijuana. <sup>5</sup>



The screenshot shows a web browser window displaying the GBIF dataset page for the Nearctic Spider Database. The browser's address bar shows the URL [www.gbif.org/dataset/82db2f8a-f762-11e1-a439-00145eb45e9a](https://www.gbif.org/dataset/82db2f8a-f762-11e1-a439-00145eb45e9a). The page has a green header with navigation links: Get data, How-to, Tools, Community, and About. Below the header, the text "OCCURRENCE DATASET | REGISTERED APRIL 3, 2007" is displayed. The main title "Nearctic Spider Database" is in large orange letters. Below the title, it says "Published by [Nearctic Spider Database](#)" and "Deleted on December 3, 2018". A navigation bar includes links for DATASET, METRICS, ACTIVITY, DOWNLOAD, and HOME PAGE. On the right, there are two buttons: "0 OCCURRENCES" and "21 CITATIONS". The main content area has a light blue background and contains a box with the text "No description available" on the left and metadata on the right. The metadata includes "Metadata last modified: December 3, 2018", "Hosted by: [Nearctic Spider Database](#)", and "Licence: [CC BY 4.0](#)". At the bottom right of the metadata box, there is a "How to cite" button and a DOI link: [DOI 10.15468/6hui3t](https://doi.org/10.15468/6hui3t).

www.gbif.org/dataset/82db2f8a-f762-11e1-a439-00145eb45e9a



Get data How-to Tools Community About

OCCURRENCE DATASET | REGISTERED APRIL 3, 2007

## Nearctic Spider Database

Published by [Nearctic Spider Database](#)

Deleted on December 3, 2018

DATASET METRICS ACTIVITY  DOWNLOAD  HOME PAGE



0 OCCURRENCES 21 CITATIONS

No description available

Metadata last modified: December 3, 2018

Hosted by: [Nearctic Spider Database](#)

Licence: [CC BY 4.0](#)

 How to cite  DOI 10.15468/6hui3t

## Reading the Small Print: GBIF Data User Agreement

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. <sup>6</sup>.

---

<sup>6</sup><https://www.gbif.org/terms/data-user> as accessed on 2025-10-13

# Data Backup Guidance of Symbiota Support Hub

Aside from regularly downloading data backup as a Darwin Core Archive, the Symbiota Support Hub promotes publishing the data to GBIF to create a secondary copy of your data <sup>7</sup>.

 [symbiota.org/portal-status-update/](https://symbiota.org/portal-status-update/)

Home

Symbiota Portals

Events ▼

Resources ▼

Th

## *1) REGULARLY DOWNLOAD DATA BACKUP FILES*

Live-managed collections should follow these [instructions](#) to download a backup copy of your data. Determining how often to download a data backup file depends on your collection's activity with respect to adding and editing records directly in the portal. For example, if your collection is actively digitizing specimen data, you might opt to download a backup file once per week; less active collections may do so less frequently. The SSH recommends that you 1) determine a backup schedule that works for your collection and then 2) create a calendar reminder (or similar) to prompt you to download a backup file according to your preferred schedule. Store your backup files in a secure location.

## *2) PUBLISH YOUR DATA TO GBIF*

Publishing your data to GBIF will create a secondary copy of your data that can be

# Data Backup Guidance of Symbiota Support Hub

## Suggested Update

Aside from regularly downloading data backup as a Darwin Core Archive, the Symbiota Support Hub promotes ~~publishing the data to~~ **registering** data with GBIF to ~~create a secondary copy~~ **promote the existence** (and use) of your data.

because ...

# Data Backup Guidance of Symbiota Support Hub

## Suggested Update

Aside from regularly downloading data backup as a Darwin Core Archive, the Symbiota Support Hub promotes ~~publishing the data to~~ **registering** data with GBIF to ~~create a secondary copy~~ **promote the existence** (and use) of your data.

because ...

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

# Why does our Digital Biodiversity Data Keep Disappearing?

~~Theory 1. People believe in the Data Fairy.~~



# Why does our Digital Biodiversity Data Keep Disappearing?

~~Theory 1. People believe in the Data Fairy.~~

Theory 2. We need better (automated) methods to bundle, publish, reference and review digital biodiversity data.

# 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: GloBI's Archive and Review of SCAN
- ▶ Example 3: Plazi's BHL Corpus

## 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?

## Introducing Digital Book Binding

In order to preserve our digital biodiversity legacy we need to systematically keep authentic transferable copies of digital biodiversity collections across digital storage media around the world. Similar to how academic libraries around the world keep physical copies of scholarly works. While methods for digitally binding biodiversity data collections exist, they have yet to be adopted by infrastructures such as iDigBio, GBIF, OBIS, etc. This lack of adoption put our data at risk for loss due to single point of failure or human error. This makes transfer, and archiving of, biodiversity data difficult at best.

# Introducing Digital Book Binding: A Few “Simple” Steps

Step 1. Reference digital data you'd like to using signed citations <sup>8</sup>.

Step 2. Describe, in a digital text file, the origin of the referenced digital data in a data bill of material (DataBoM).

Step 3. Publish the data bill of material (DataBoM).

Step 4. Ensure that bill of material **and the cited data** are stored in across independent locations and storage media.

Step 5. Use the signed citation of the data bill of material (DataBoM) in your research.

Step 6. Continuously monitor the availability of the DataBoM and the associated data.

---

<sup>8</sup>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 <sup>9</sup> command

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

---

<sup>9</sup><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 <sup>10</sup> and Hash URIs <sup>11</sup> 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...> .
```

---

<sup>10</sup><https://www.w3.org/TR/prov-o/>

<sup>11</sup>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 with signature X
2. list data signatures in DataBoM
3. get data associated with cited 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

Details

Comments

Linked Resources



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

## Extra Credit: Finding the Last Record

What bash script would find the last DwC record associated with this DataBoM ?

- ▶ Hint: lizards have one, but humans don't.

## Extra Credit: Finding the Last Record

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

... associates with an occurrence of fungus *Absidia corymbifera* .

# 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 combination 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>

# Why does our Digital Biodiversity Data Keep Disappearing?

~~Theory 1. People believe in the Data Fairy.~~

~~Theory 2. We need better (automated) methods to bundle, publish, reference and review digital biodiversity data.~~

Theory 3. When using digital data in publications, we need to increase support for (and use of) textual Data Bill of Materials (DataBoM) with embedded **signed** data citations so that digital data can be bundled, cited and securely transferred regardless of their data format or location.



# How to keep our Digital Biodiversity Data around?

Suggest to:

1. cite your *original* source data
2. use **signed** data citations
3. publish the data you produce
4. republish the *original* data you use

or... come up with another method to digitally bind datasets.

## And Remember



Thank you!

Made possible (in part) by 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](mailto:jhpoelen@jhpoelen.nl)

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