Fairy Tales and Data Reuse

Jorrit H. Poelen

2025-11-19

Presented at the 17-19 Nov 2025 kick-off meeting in Freising, 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. (2025) Fairy Tales and Digital Research Data. hash://md5/74cabb19c6dcf3e2eea27a38acf4fb76 Zenodo. https://doi.org/10.5281/zenodo.17625448

Cite As

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

License

Guiding Question

How do *you* ensure your data is reused?

Digital Data are hard to reuse

[. . .] In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship'[1] *were published in Scientific Data. The authors intended to provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles* ***emphasise machine-actionability*** *(i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the* ***increase in volume, complexity, and creation speed of data****. [. . .]* [2]

[1] Wilkinson, et al. 2016, Sci Data doi:10.1038/sdata.2016.18

[2] Accessed on 2025-11-19 at https://www.go-fair.org/fair-principles/.

Biodiversity Data are hard to reuse

[...] "***The various books and journals of ornithology and entomology are like a row of beehives containing an immense amount of valuable honey, which has been stored up in separate cells by the bees that made it.*** *The advantage, and at the same time the difficulty, of ecological work is that it attempts to provide conceptions which can link up into some complete scheme the colossal store of facts about natural history which has accumulated up to date in this rather haphazard manner. [...] Until more organised information about the subject is available, it is only possible to give a few instances of some of the more clearcut niches which happen to have been worked out. [...]*" [3]

[3] Charles Elton, 1927. Animal Ecology. pp 65-66. doi:10.5962/bhl.title.7435

Biodiversity Data are hard to reuse

[. . .] "The various books and journals of ornithology and entomology are like a row of beehives containing an immense amount of valuable honey, which has been stored up in separate cells by the bees that made it. ***The advantage, and at the same time the difficulty, of ecological work is that it attempts to provide conceptions which can link up into some complete scheme the colossal store of facts about natural history which has accumulated up to date in this rather haphazard manner.*** *[. . .] Until more organised information about the subject is available, it is only possible to give a few instances of some of the more clearcut niches which happen to have been worked out. [. . .]"* [4]

[4] Charles Elton, 1927. Animal Ecology. pp 65-66. doi:10.5962/bhl.title.7435

Biodiversity Data are hard to reuse

[. . .] "The various books and journals of ornithology and entomology are like a row of beehives containing an immense amount of valuable honey, which has been stored up in separate cells by the bees that made it. The advantage, and at the same time the difficulty, of ecological work is that it attempts to provide conceptions which can link up into some complete scheme the colossal store of facts about natural history which has accumulated up to date in this rather haphazard manner. [. . .] ***Until more organised information about the subject is available, it is only possible to give a few instances of some of the more clearcut niches which happen to have been worked out.*** *[. . .]"* [5]

[5] Charles Elton, 1927. Animal Ecology. pp 65-66. doi:10.5962/bhl.title.7435

Who or what keeps our Digital Biodiversity Data reusable?

a) **the FAIR Data Fairy.**

b) Awareness of the complexity of reusing digital data.

c) Common sense data reuse and review practices.

Who or what keeps our Digital Biodiversity Data reusable?

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

b) Awareness of the complexity of reusing digital data.

c) Common sense data reuse and review practices.

Is GBIF our FAIR 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. [6].

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

Is GBIF our FAIR 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. [7].

No!

GBIF provides a framework in which standardized biodiversity datasets can be registered, reviewed, and queried. However, you are responsible for ensuring that the data continues to be fit for reuse.

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

Is Zenodo our FAIR 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. [...] [8].

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

Is Zenodo our FAIR 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. [...] [9].

No!

Zenodo provides a framework in which digital works can be deposited, found, and retrieved. However, you are responsible for ensuring that the data continues to be fit for reuse.

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

FAIR Data Fairy is a Cousin of the Poop Fairy.

FAIR Data Fairy is a Cousin of the Poop Fairy.

Recent Reminders of FAIR Data Fairy Absence

Scholarly data...

- cannot be found because it is left unpublished on a personal laptop or desk drawer
- cannot be accessed due to "data available on request" of retired/deceased author
- is lost due to malicious activity (e.g., hacking, ransonware) or human error (oops!)
- is published using custom schemas instead of schemas used by others
- is published in proprietary or machine unfriendly file formats (e.g., xlsx, docx, pdf)
- is riddled with data inconsistencies and typos
- cannot be easily reused/linked/integrated with other datasets/projects
- are effectively excluded during peer review process due to time constraints

Who or what keeps our Digital Biodiversity Data reusable?

a) ~~the FAIR Data Fairy~~

b) Awareness of the complexity of reusing digital data.

c) Common sense data reuse and review practices.

Guiding Questions

How do *you* reuse/review your own (past, present, future) digital research data?

How do *you* reuse/review third party (past, present, future) digital research data?

Who or what keeps our Digital Biodiversity Data reusable?

a) ~~the FAIR Data Fairy~~

b) Awareness of the complexity of reusing digital data.

c) Common sense data reuse and review practices.

Peer Review for Digital Biodiversity Data

- What is so neat about peer review?
- How to review digital data?

The Neat Thing about Peer Review

~~Even before~~ Mainly after the invention of the book press, ~~books and scrolls~~ **scholarly journals and academic societies**[10] have been pretty successful in transferring **scientific** knowledge across generations and around the world.

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

~~Books~~ **Printed scholarly journals** are wireless, their content cannot be easily altered remotely, changes can be detected (ripped out pages), and they need no power to operate. (note to self: archive digital only journals)

Typically, scholarly journals employ a peer review process to select scholarly works of interest and increase their quality through review cycles and discourse.

[10] First journal "Philosophical Transactions" was launched in 1665 according to https://royalsocietypublishing.org/journal/rstl accessed on 2025-11-19.

The Neat Thing about Physical Books

~~Even before~~ Mainly after the invention of the book press, ~~books and scrolls~~ **scholarly journals and academic societies**[11] have been pretty successful in transferring **scientific** knowledge across generations and around the world.

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

~~Books~~ **Printed scholarly journals** are wireless, their content cannot be easily altered remotely, changes can be detected (ripped out pages), and they need no power to operate. (note to self: archive digital only journals)

Typically, scholarly journals employ a peer review process to select scholarly works of interest and increase their quality through review cycles and discourse.

Idea . . . what if we treat digital data more like a peer-reviewed publication instead of some haphazardly organized data appendix?

[11] First journal "Philosophical Transactions" was launched in 1665 according to https://royalsocietypublishing.org/journal/rstl accessed on 2025-11-19.

Steps to Reviewing Digital Data

To enable the discovery of existing species interaction datasets, Global Biotic Interactions (GloBI) continuously tracks existing datasets and integrates the discovered interaction records. These integrated interaction records form the basis of the GloBI's interpreted interaction data. [12]

Step 1. *automatically* discover species interaction datasets

Step 2. *automatically access* up-to-date datasets and versioned them

Step 3. *automatically parse* data and link/integrate with other datasets (e.g., taxonomic resources)

Step 4. *automatically produce* derived data products and data services

Step 5. *automatically publish* derived datasets and associated GloBI data reviews

Step 6. *semi-manual* solicit feedback and encourage discussion through open access and

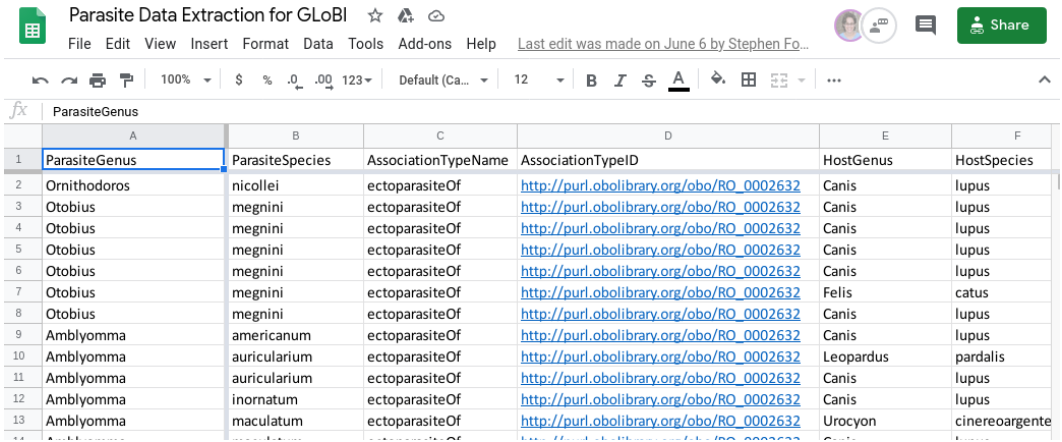
Step 7. repeat continuously until (not yet planned) retirement

[12] https://globalbioticinteractions.org/process accessed on 2025-11-19

Steps to Reviewing Digital Data: Example

Figure 1: A spinose ear tick (Otobius megnini) a mammalian parasites (Lindström, A. 2017). from: Fowler, S. 2020. Extracting Parasite Interaction Data from a Scientific Paper using Google Sheets. Accessed on 2025-11-19 on https://www.globalbioticinteractions.org/2020/08/25/extracting-parasite-interaction-data/

Steps to Reviewing Digital Data: Example



	A	B	C	D	E	F
1	ParasiteGenus	ParasiteSpecies	AssociationTypeName	AssociationTypeID	HostGenus	HostSpecies
2	Ornithodoros	nicollei	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
3	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
4	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
5	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
6	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
7	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Felis	catus
8	Otobius	megnini	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
9	Amblyomma	americanum	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
10	Amblyomma	auricularium	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Leopardus	pardalis
11	Amblyomma	auricularium	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
12	Amblyomma	inornatum	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Canis	lupus
13	Amblyomma	maculatum	ectoparasiteOf	http://purl.obolibrary.org/obo/RO_0002632	Urocyon	cinereoargente

Figure 2: An example of the Google Sheets spreadsheet used for transcribing data. GloBI accessed the transcribed data set through open source scripts accessible on GitHub. Once every other day the data was indexed into GloBI data products. (Fowler, S. 2020)

Steps to Reviewing Digital Data: Example

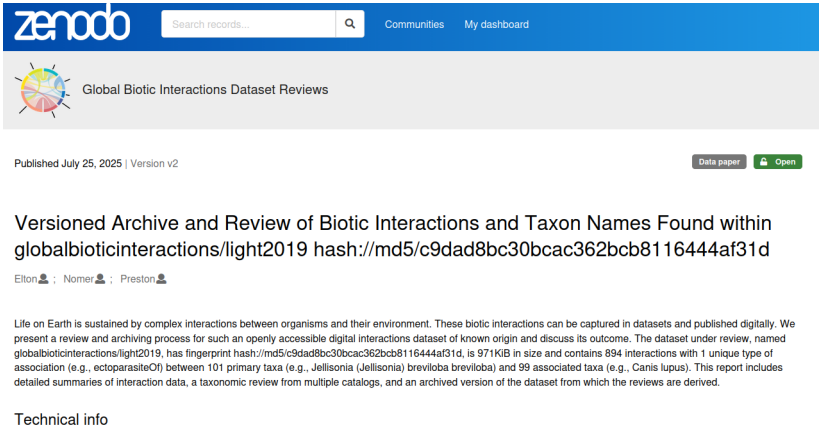


Figure 3: Elton, Nomer, & Preston. (2025). Versioned Archive and Review of Biotic Interactions and Taxon Names Found within globalbioticinteractions/light2019 hash://md5/c9dad8bc30bcac362bcb8116444af31d. Zenodo. doi:10.5281/zenodo.16416949

Steps to Reviewing Digital Data: Example

Figure 4: GloBI interactive search results as accessed on 2025-11-19 via https://www.globalbioticinteractions.org

Steps to Reviewing Digital Data: Example

Figure 5: GloBI interactive search results as accessed on 2025-11-19 via https://www.globalbioticinteractions.org

Steps to Reviewing Digital Data: Example

Step 1. use Elton [13] to discover dataset metadata at https://github.com/globalbioticinteractions/light2019

Step 2. use Elton to access and version csv file available through Google Sheet 1Fo. . . DxY

Step 3. use Elton and Nomer to align taxonomic names with NCBI taxonomy, Catalogue of Life etc.

Step 4. use Elton, Nomer, pandoc to produce derived datasets and a data review.

Step 5. use Preston to *publish* derived datasets and associated GloBI data review to Zenodo.

Step 6. *semi-manual* solicit feedback and encourage discussion through email, meetings and open data platforms

Step 7. repeat continuously until (not yet planned) retirement

[13] For review robots description see: https://globalbioticinteractions.org/process

Who or what keeps our Digital Biodiversity Data around?

a) ~~the FAIR Data Fairy~~

b) Awareness of the complexity of reusing digital data.

c) Common sense data reuse and review practices.

Guiding Question

What is *one* of *your* experiences that comes to mind in which digital data was unfindable, riddled with errors or otherwise hard to reuse? Share your horror story!

And Remember

Thank you!

Made possible (in part) by Horizon Europe's 101219108

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 Question

How do *you* ensure your data is reused?