Title: To connect is to preserve: on frugal data integration and preservation solutions

Author(s): Jorrit H Poelen

Affiliation: , Independent, Global Biotic Interactions, 400 Perkins St Apt 104, Oakland, CA, 94610, USA

Abstract: The deluge of digital biodiversity datasets unleashed through institutional, national and global infrastructures brings up an inconvenient truth: internet-connected infrastructures are in a constant state of flux while preservation and integration of digital knowledge are often afterthoughts. Rather than taking digital amnesia for granted, we examine examples of durable and frugal digital data preservation and integration methods. Examples include tracking external datasets, creating verifiable data citations, cross-publishing and cross-linking datasets, reproducing data-integration processes, and distributing large data archives across poor, or nonexistent, internet connections. Topics include cryptographic hashes, Provenance Ontology, content-addressed storage, Unix philosophy, and offline first design as applied in projects like Preston (https://preston.guoda.bio) and Global Biotic Interactions (https://globalbioticinteractions.org). The examples are then related to best practices applied by proven knowledge-preservation experts: librarians and curators.

Keywords: biodiversity data, data preservation, data integration, biodiversity informatics

To Connect Is to Preserve

On Frugal Data Integration and Preservation

Jorrit H. Poelen

Independent Biodiversity Informatics
/ Software Engineer
jhpoelen@xs4all.nl

How are we going to keep our digital datasets alive and easy to use for the next 50 years?

"[...] digital formats and the workflows that go with them change all the time, so collections need to stay on their toes if they don't want to lose data and the ideas that go with them."

- Kate Webbink (2014)
Information Systems Specialist,
Field Museum

"Just because you don't have the budget, time and staff, it doesn't mean you should not just try and present it, and maybe somebody has some money for you somewhere."

- Janeen Jones (2018)
Assistant Collections Manager, Invertebrates & Database
Admin for Collections, Field Museum

What is the state of our digital biodiversity datasets?

Publish and Retire Scenario

A scientist has spent 20 years collecting, digitizing and analyzing parasitic lice (Geomydoecus sp.) of gophers (Geomyidae) occurrences in Texas.

He makes his data available using his data portal and registers with biodiversity networks.

After his retirement, his publications remain accessible through traditional scientific journals.

However, he is unable to maintain his data portal and 20 years of digital datasets are lost forever.

The "Ooops" Scenario

After a collection manager registers a vertebrate dataset with many biodiversity data networks, a researcher uses that dataset to study vertebrates.

Some time later, the collection manager mistakenly deletes all Chordata records.

Subsequently, the researcher does a follow-up study and is no longer able to find his source datasets.

The data is lost forever because an archive of older datasets is not available.

What is the state of our digital biodiversity datasets?

Internet Terms Refresher

Internet Terms Refresher

IP Addresses ~ a "physical" address on the internet, or the network location (e.g., 35.184.133.11). Centrally administered by the Internet Assigned Numbers Authority (IANA).

Domain Name ~ human-readable shortcut to one (or more) IP addresses (e.g., spnhcchicago2019.com points to 35.184.133.11). Centrally administered by IANA.

Uniform Resource Locator (URL) ~ "refers to the subset of URIs that, in addition to identifying a resource, provide a means of locating the resource by describing its primary access mechanism"* (e.g., **its *network* "location"**)

* https://www.ietf.org/rfc/rfc3986.txt

Internet Terms Refresher - Examples

Valid URLs

https://doi.org/10.1016/j.ecoinf.2014.08.005

http://arctos.database.museum/guid/MVZ:Bird:180448

Invalid URLs

d2a7ca86-8181-11e9-9c58-fb664cf4fb3e

hash://sha256/3102dae4b68cebe40337730312fcb612297b8928547267e8b3d1ee6002b2d683

What is the state of our digital biodiversity datasets?

What is the state of our digital biodiversity datasets?

Hypothesis:

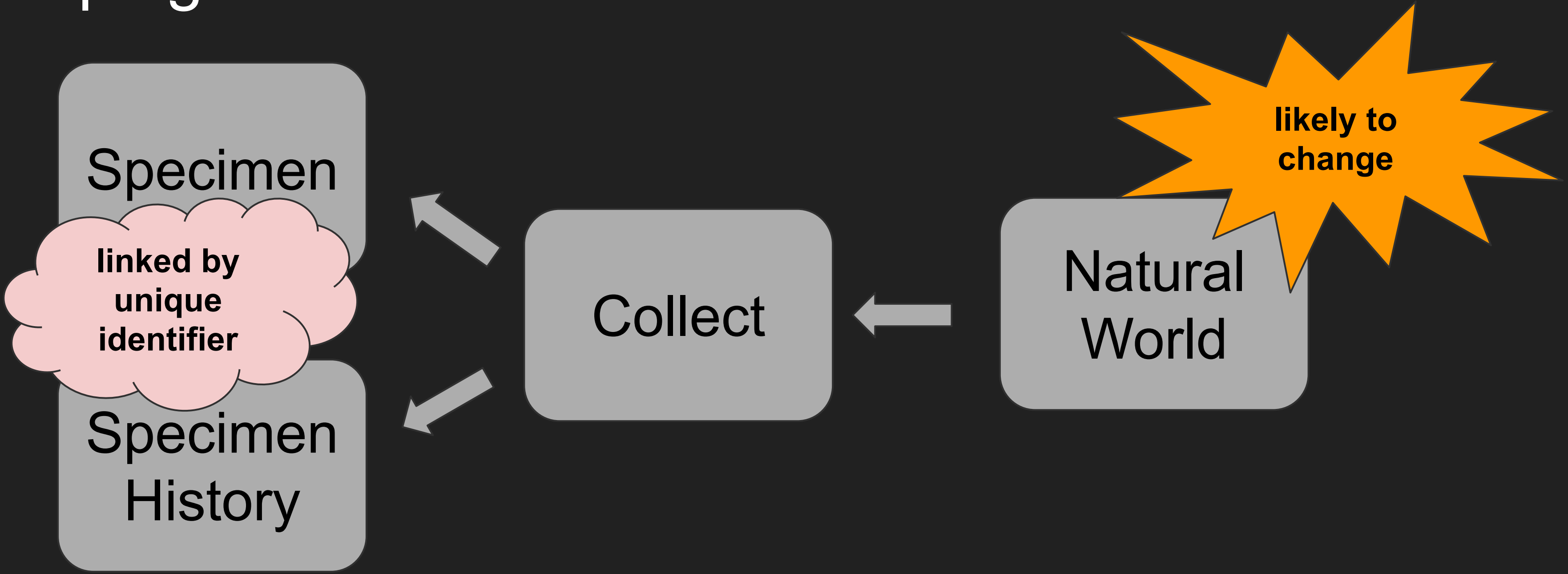
Digital biodiversity datasets are published, accessed and linked using methods unsuitable for data archiving. This causes digital datasets to go offline (link rot) or uncontrollably change (content drift).

URLs, domain names and IP addresses are designed to communicate information, not to archive it.

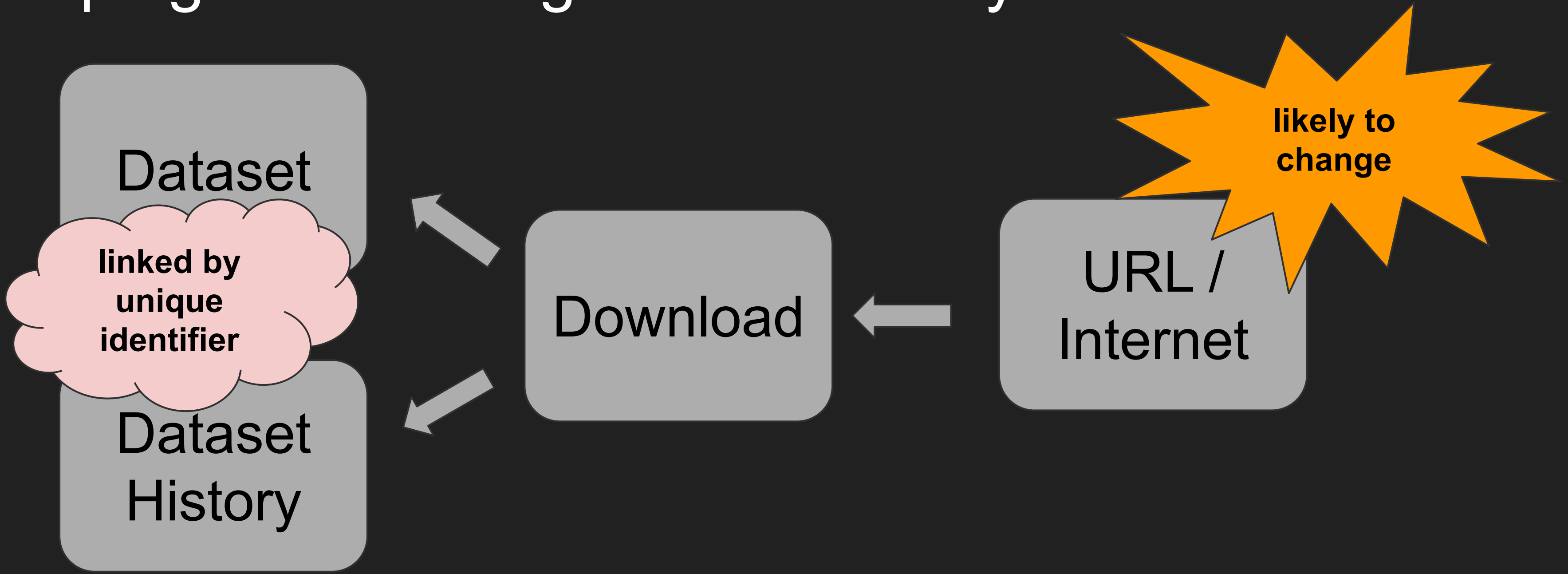
A method:

Use frugal tools to continuously monitor and archive digital datasets from major biodiversity networks in order to record their availability.

Keeping Track of the Natural World



Keeping Track of Digital Biodiversity Datasets



Unique Specimen Identifiers

Collect specimen and use unique identifiers that do not rely on where the specimen happens to be stored or when it was collected.

Instead of:

Field Museum, Vertebrate Collection, specimen 352

Use an algorithmically generated, globally unique identifier, such as:

d2a7ca86-8181-11e9-9c58-fb664cf4fb3e

Unique ~~Specimen~~ Data Identifiers

Special algorithms (e.g., sha256*) exist that can **generate a unique, fixed-length identifier for any digital data from the data itself.**

Examples:

sha256 ("hello") =
5891b5b522d5df086d0ff0b110fbd9d21bb4fc7163af34d08286a2e846f6be03

sha256 ("Hello") =
66a045b452102c59d840ec097d59d9467e13a3f34f6494e539ffd32c1bb35f18

* Device for and method of one-way cryptographic hashing. US Patent 6829355.

Keeping Track of the Original Specimen and Associated Information

What not to do:

Record the collection date, location, etc. and toss the original specimen.

Instead:

Record the collection date, location, etc. and **keep the specimen** and **relate the specimen to the record**.

Keeping Track of the Original ~~Specimen~~ Data and Associated Information

Use flexible, expressive and easy-to-aggregate machine-readable "English"* in the form of short sentences that looks like:

<subject> <verb> <object> .

* Resource Definition Language (RDF)

Keeping Track of the Original ~~Specimen~~ Data and Associated Information

For example:

https://example.com/data has version dataset X

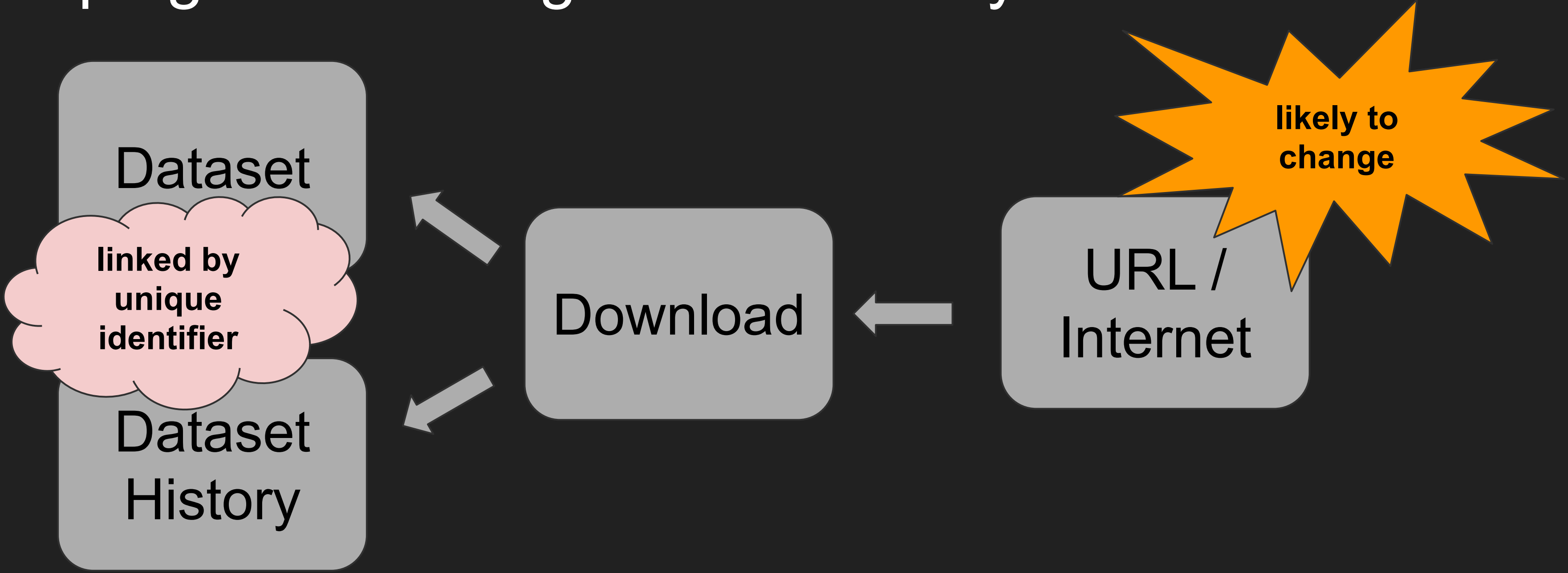
dataset X was generated on 27 May 2019

dataset X was generated by a download process Z

Download process Z was started by Jane Doe

Please see https://preston.guoda.bio for explicit examples using RDF, PROV-O and PAV.

Keeping Track of Digital Biodiversity Datasets



Takeaways

URLs, domain names and IP addresses are designed to transfer data, not to keep data permanently accessible.

Link rot (missing data) and **content drift** (changing data) is likely occurring in biodiversity networks just like anywhere else, and I am currently monitoring both to quantify.

To keep track of our digital datasets, we can apply best practices of natural history collections and help preserve our digital heritage.

We need to help establish a distributed and trusted network of biodiversity datasets.

Related Initiatives / Technologies / Keywords

https://preston.guoda.bio - a biodiversity dataset tracker prototype

https://hash-archive.org - a registry of content hashes

https://archive.org - the Internet Archive

Resource Definition Format (RDF)

Provenance Ontologies

Content-based addressing

Decentralized architectures

Acknowledgments

Thanks to **Matt Collins, José A. B. Fortes, Michael Elliot, Icaro Alzuru** and many others.

With support from Advanced Computing and Information Systems (**ACIS**) Laboratory, University of Florida, Gainesville.

This work is funded in part by grant NSF OAC 1839201 from the National Science Foundation.

How are we going to keep our digital datasets alive and easy to use for the next 50 years?