Conference

4th Annual Digital Data in Biodiversity Research Conference June 1-3 2020

Title

Reliable dataset identifiers are essential building blocks for reproducible research

Type

Oral presentation, Discussion

Authors

Jorrit Poelen, Michael Elliott, Jose Fortes

Abstract

Despite increased use of digital biodiversity data in research, reliable methods to identify datasets are not widely adopted. While commonly used location-based dataset identifiers such as URLs help to easily download data today, additional identification schemes are needed to ensure long term access to datasets.

We propose to augment existing location- and DOI-based identification schemes with cryptographic content-based identifiers. These content-based identifiers can be calculated from the datasets themselves using available cryptographic hashing algorithms (e.g., sha256). These algorithms take only the digital content as input to generate a unique identifier without needing a centralized identification administration. The use of content-based identifiers is not new, but a re-application of change management techniques used in the popular version control system "git".

We show how content-based identifiers can be used to version datasets, to track the dataset locations, to monitor their reliability, and to efficiently detect dataset changes. We discuss the results of using our approach on datasets registered in GBIF and iDigBio from Sept 2018 to May 2020. Also, we propose how reliable, decentralized, dataset indexing and archiving systems can be devised. Lastly, we outline a modification to existing data citation practices to help work towards more reproducible and reusable research workflows.

Access this page via:
https://tinyurl.com/reliable-data

Reliable dataset identifiers are essential building blocks for reproducible research

1 June 2020, 4th Annual Digital Data in Biodiversity Research, see "Reliable dataset identifiers are essential building blocks for reproducible research" .

Jorrit Poelen, Ronin Institute and independent software engineer;

Michael Elliott, Advanced Computing and Information Systems Laboratory (ACIS) Department of Electrical and Computer Engineering, University of Florida;

Jose Fortes, Advanced Computing and Information Systems Laboratory (ACIS) Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL

This work is funded in part by grant NSF OAC 1839201 "EAGER: Towards the Web of Biodiversity Knowledge: Understanding Data Connectedness to Improve Identifier Practices" from the **National Science Foundation**.

Reliable Dataset References

Michael Elliott, Jorrit H. Poelen, José A. B. Fortes

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

What does an unreliable reference look like?

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- An example reference using a URL

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

Get data How-to Tools Community About

https://doi.org/10.15468/aomfnb

OCCURRENCE DATASET | REGISTERED SEPTEMBER 16, 2010

EOD - eBird Observation Dataset

Published by Cornell Lab of Ornithology

Tim Levatich • Francisco Padilla • Jeff Gerbracht

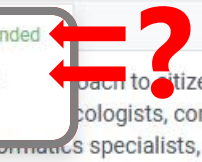
DATASET METRICS ACTIVITY DOWNLOAD

561,852,542 OCCURRENCES 104 CITATIONS

GBIF annotated archive Recommended

Source archive Darwin Core Archive

GBIF annotated metadata EML



eBird is a ...pach to citizen science by developing cooperative partnerships among ex... ...cologists, conservation biologists, quantitative ecologists, statisticians, computer scientists, GIS and informatics specialists, application developers, and data administrators. Managed by the Cornell Lab of Ornithology eBird's goal is to increase data quantity through participant recruitment and engagement globally, but also... More

Metadata last modified: March 27, 2019

Data last changed: March 30, 2019

Hosted by: Cornell Lab of Ornithology

License: CC0 1.0

How to cite DOI 10.15468/aomfnb

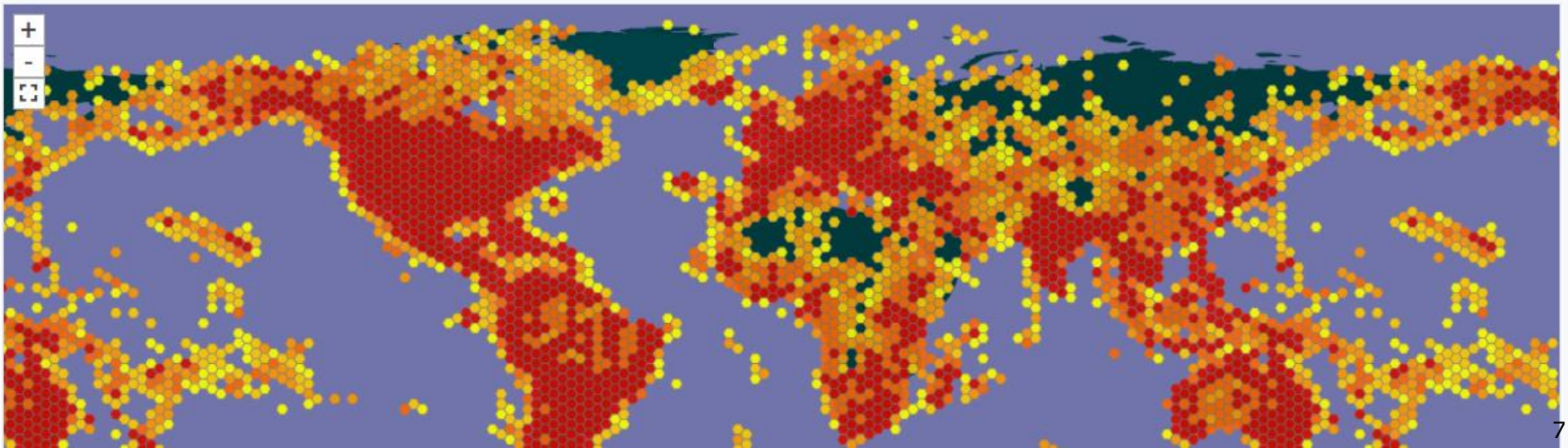
561,852,542 Occurrences

100% With taxon match

99.9% With coordinates

100% With year

561,766,080 GEOREFERENCED RECORDS



Get data How-to Tools Community About

https://doi.org/10.15468/aomfnb

OCCURRENCE DATASET | REGISTERED SEPTEMBER 16, 2010

EOD - eBird Observation Dataset

Published by Cornell Lab of Ornithology

Tim Levatich • Francisco Padilla • Jeff Gerbracht

DATASET METRICS ACTIVITY DOWNLOAD

561,852,542 OCCURRENCES 104 CITATIONS

GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a citizen science by developing cooperative partnerships among ex conservation biologists, quantitative ecologists, statisticians, computer scientists, GIS and informatics speci ts, application developers, and data administrators. Managed by the Cornell Lab of Ornithology eBird's goal is to rease data quantity through participant recruitment and engagement globally, but also... More

Metadata last modified: March 27, 2019
Data last changed: March 30, 2019
Hosted by: Cornell Lab of Ornithology
License: CC0 1.0

561,8
Occur

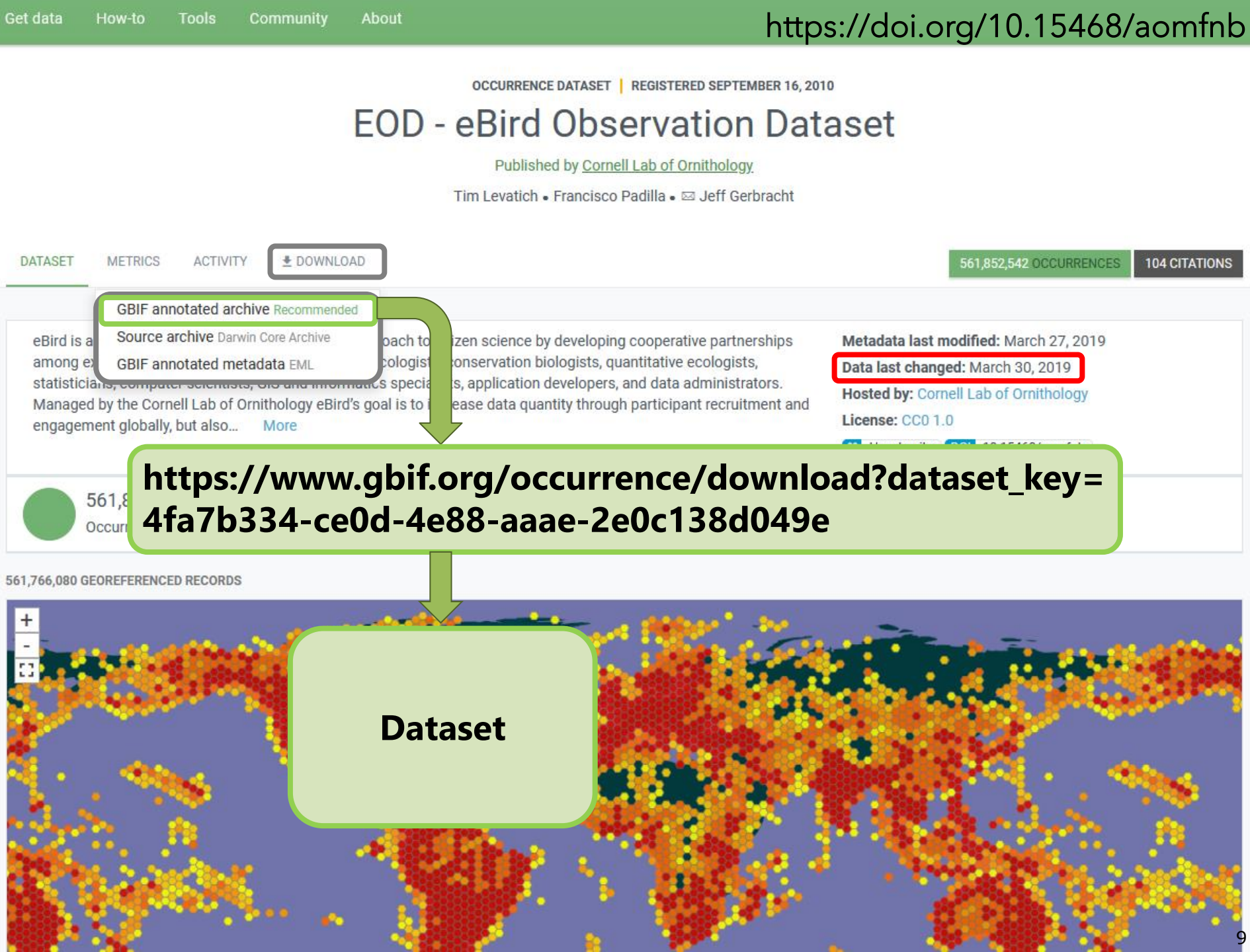
http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip

561,766,080 GEOREFERENCED RECORDS

COX

ebirddata.ornith.cornell.edu cannot be found

Suggested searches


Get data
How-to
Tools
Community
About
https://doi.org/10.15468/aomfnb
OCCURRENCE DATASET | REGISTERED SEPTEMBER 16, 2010
EOD - eBird Observation Dataset
Published by Cornell Lab of Ornithology
Tim Levatich • Francisco Padilla • Jeff Gerbracht
DATASET
METRICS
ACTIVITY
DOWNLOAD
561,852,542 OCCURRENCES
104 CITATIONS
GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML
Metadata last modified: March 27, 2019
Data last changed: March 30, 2019
Hosted by: Cornell Lab of Ornithology
License: CC0 1.0
https://www.gbif.org/occurrence/download?dataset_key=4fa7b334-ce0d-4e88-aaae-2e0c138d049e
561,766,080 GEOREFERENCED RECORDS
Dataset

Get data How-to Tools Community About

https://doi.org/10.15468/aomfnb

OCCURRENCE DATASET | REGISTERED SEPTEMBER 16, 2010

EOD - eBird Observation Dataset

Published by Cornell Lab of Ornithology

Tim Levatich • Francisco Padilla • Jeff Gerbracht

DATASET METRICS ACTIVITY DOWNLOAD

561,852,542 OCCURRENCES 104 CITATIONS

GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a ...oach to ...izen science by developing cooperative partnerships among ex... ...cologist... ...conservation biologists, quantitative ecologists, statisticians, computer scientists, GIS and informatics specia... ...s, application developers, and data administrators. Managed by the Cornell Lab of Ornithology eBird's goal is to i... ...ease data quantity through participant recruitment and eng...

Metadata last modified: March 27, 2019
Data last changed: March 30, 2019
Hosted by: Cornell Lab of Ornithology

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

561,766,...

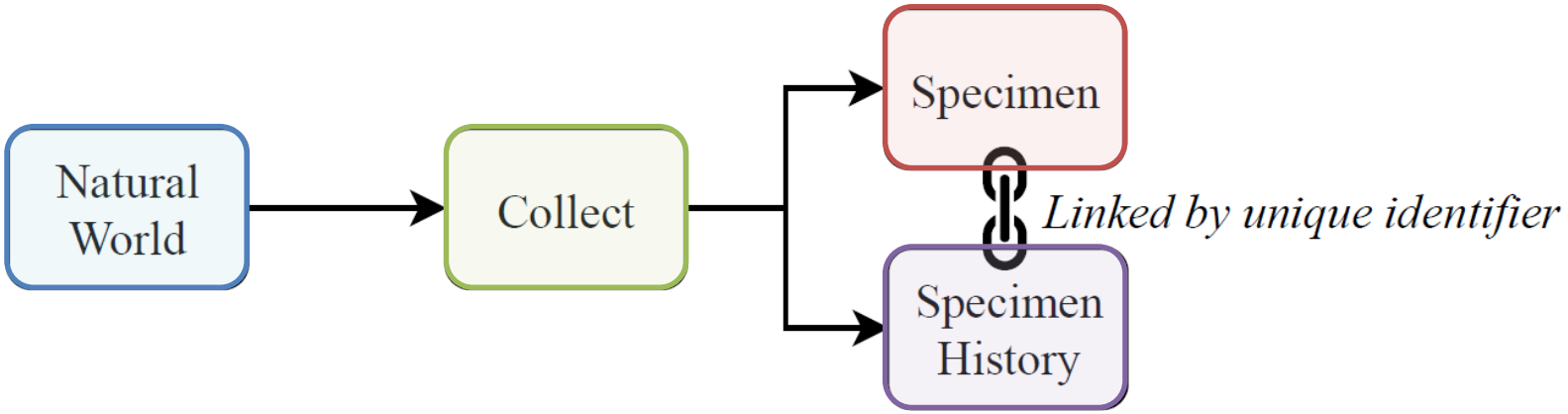
Dataset

A reference is **reliable** if it both

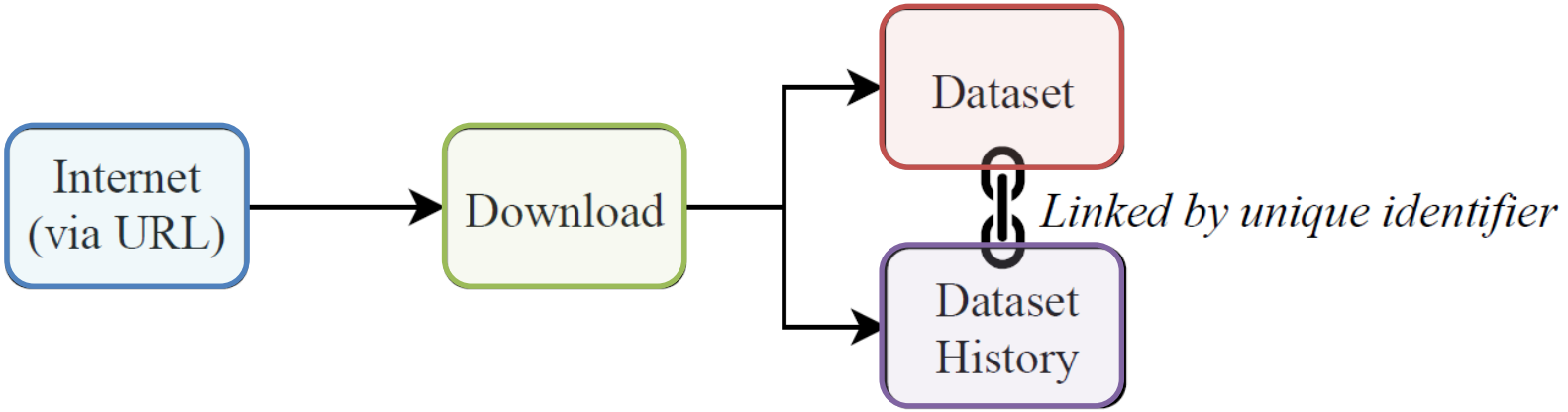
- Allows continued access to what was referenced
- Only identifies what was referenced

Location-based identifiers, such as URLs, tend to be unreliable

How can we reliably reference datasets served at unreliable URLs?



(a) Physical specimen collection



(b) Digital data collection

- Content-based identifiers can be used to identify datasets
- Cryptographic hash functions can produce unique content-based identifiers for digital datasets

Content-based identifier for the 2017 eBird dataset:

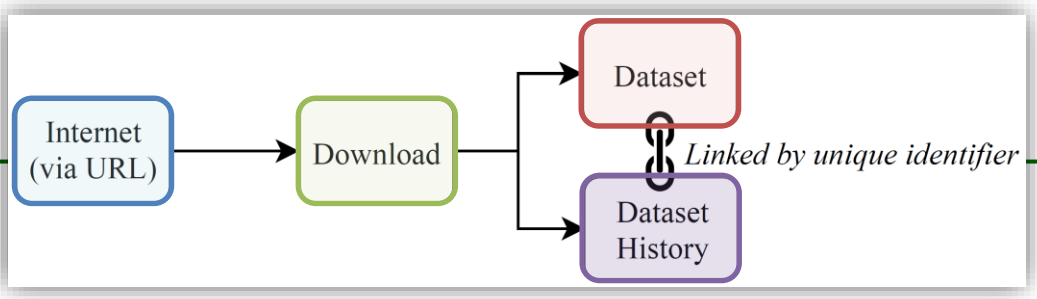
hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c

Content-based identifier for the 2019 eBird dataset:

hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA



http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip

Download at 1:25 AM on September 2, 2018

hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c

hash://sha256/c253a5311a20c2fc082bf9bac87a1ec5eb6e4e51ff936e7be20c29c8e77dee55

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

- Our example eBird reference:

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2018-09-02.

- Add a content-based identifier

Levatich T, Padilla F (2017). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset **hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c** accessed at https://doi.org/10.15468/aomfnb via GBIF.org on 2018-09-02

Hash Archive (beta)

URL or hash: hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c Lookup

Sources for hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c

- Search for this hash on Google
- Search for this hash on DuckDuckGo
- Search for this block on IPFS
- Check this hash on VirusTotal
- Other useful sources...?

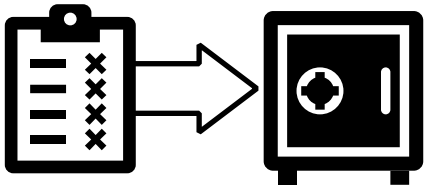
Active as of May 27th, 2020
https://zenodo.org/record/3858251/files/dwca-1.0.zip?download=1[^]

Active as of May 27th, 2020
https://archive.org/download/biodiversity-dataset-archives/data.zip/data/29/d3/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c[^]

Active as of May 26th, 2020
https://deeplinker.bio/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c[^]

Active as of February 12nd, 2020
https://archive.org/download/biodiversity-dataset-archives/data.zip/data%2F29%2Fd3%2F29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c[^]

Obsolete; last seen March 9th, 2019
http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip[^]

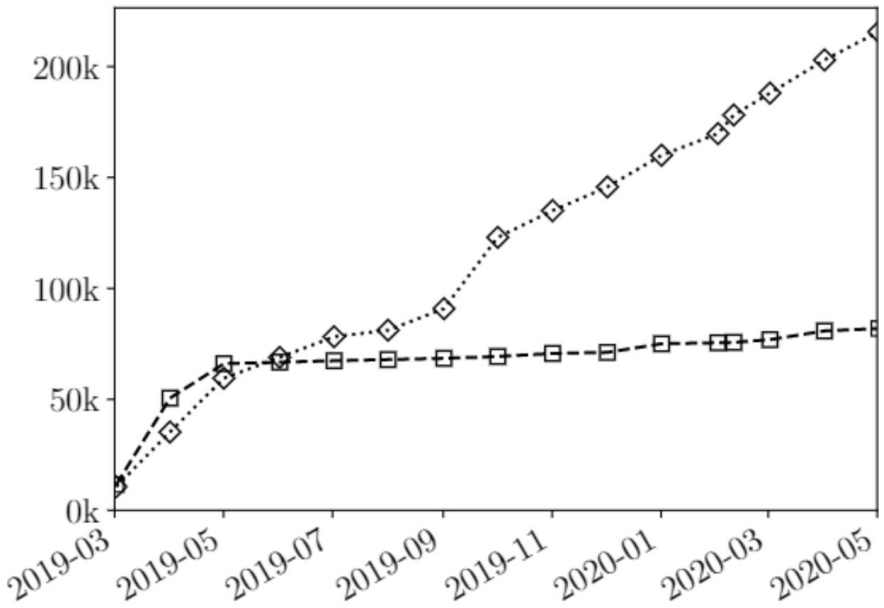


ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

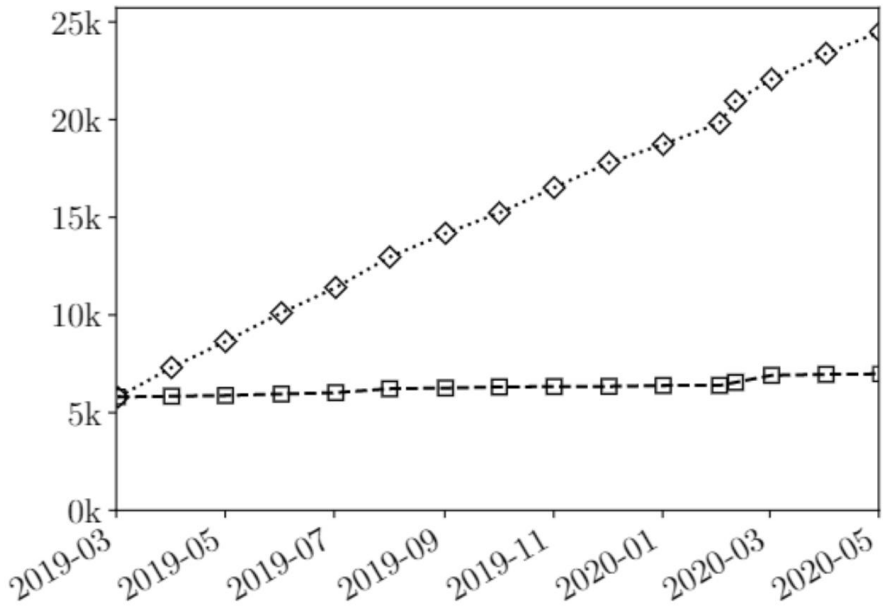
UF UNIVERSITY of FLORIDA

- In order to have reliable references,
  1) Datasets must be addressable and retrievable using content-based identifiers (rather than, e.g., URLs)
  2) Datasets that could be used in the future should be archived
  3) Openly accessible registries must exist to lookup locations for datasets referenced by content-based identifiers
  4) Agents must exist to collect datasets, record their provenance (i.e., history record), deposit them into archives, and share their new locations with public registries

- Why are reliable references important?
  - The reliability of URLs as references decreases over time



(c) GBIF

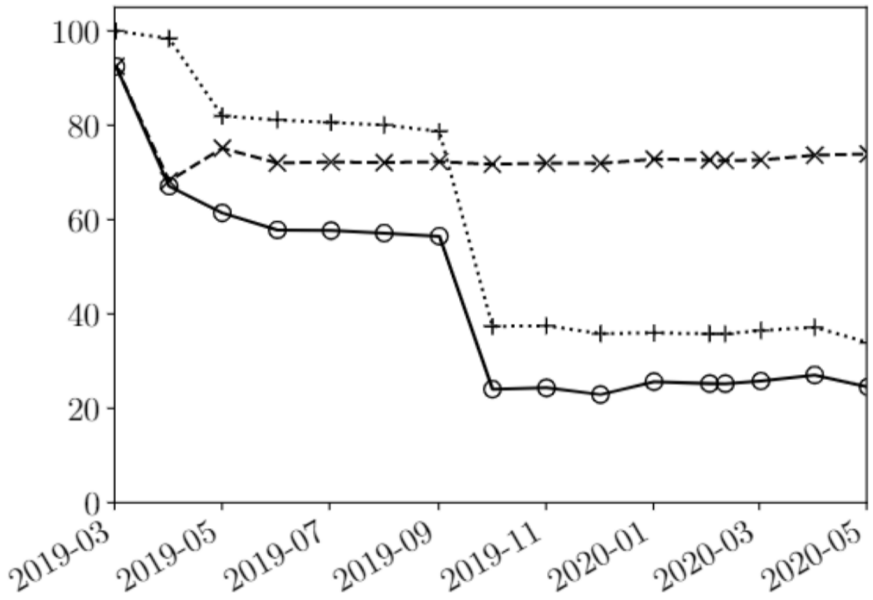


(d) iDigBio

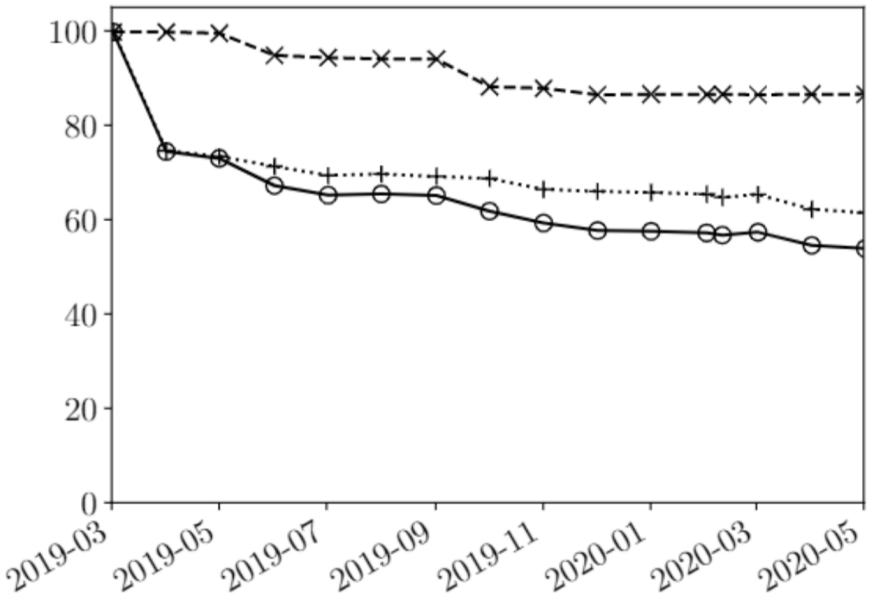
Total URLs
Total Contents

(Datasets)

- Why are reliable references important?
  - The reliability of URLs as references decreases over time



(c) GBIF



(d) iDigBio

Percent Responsive
Percent Stable
Percent Reliable

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

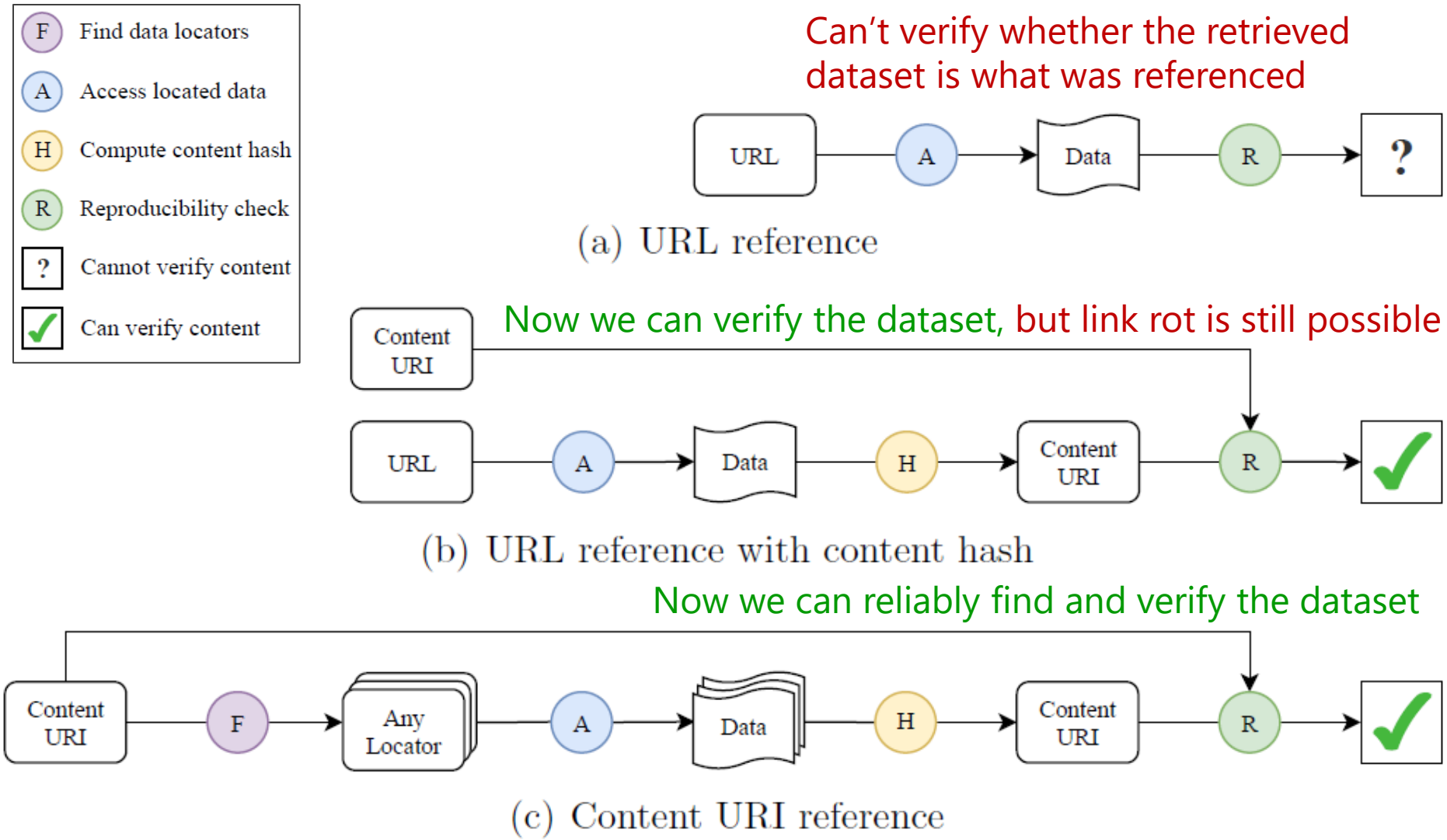
Thanks for Listening!

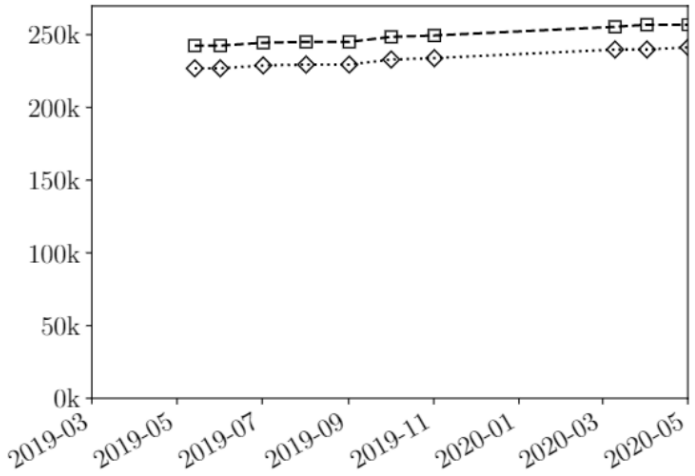
Based on our paper, "Toward Reliable Biodiversity Dataset References" (Preprint: https://ecoevorxiv.org/mysfp)

The research was funded in part by a grant (NSF OAC 1839201) from the National Science Foundation and the AT&T Foundation.

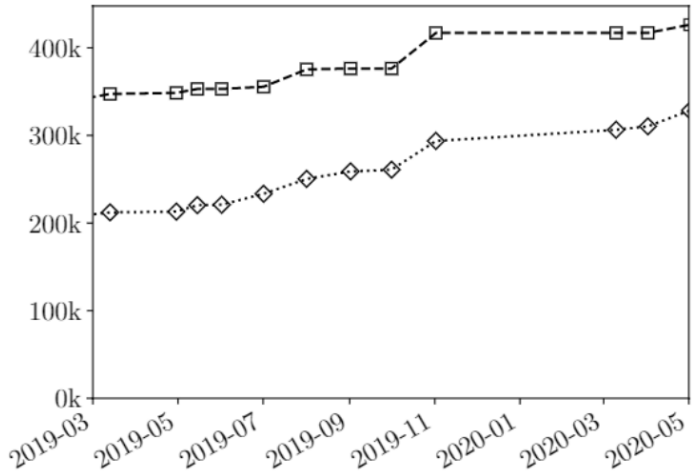
ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

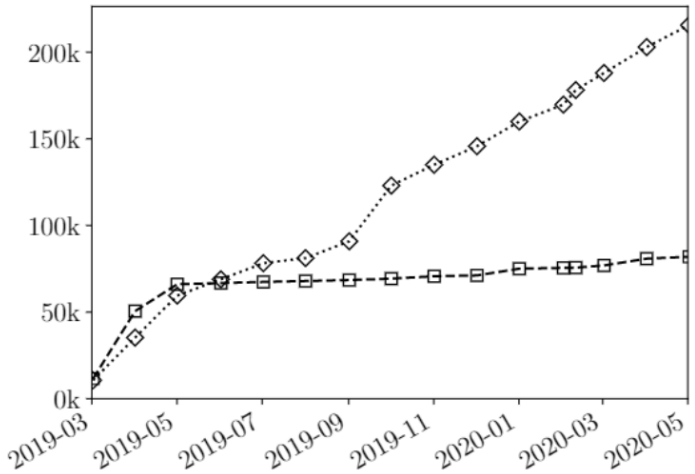

F Find data locators
A Access located data
H Compute content hash
R Reproducibility check
? Cannot verify content
Can verify content
Can't verify whether the retrieved dataset is what was referenced
URL
A
Data
R
?
(a) URL reference
Content URI
Now we can verify the dataset, but link rot is still possible
URL
A
Data
H
Content URI
R
(b) URL reference with content hash
Now we can reliably find and verify the dataset
Content URI
F
Any Locator
A
Data
H
Content URI
R
(c) Content URI reference



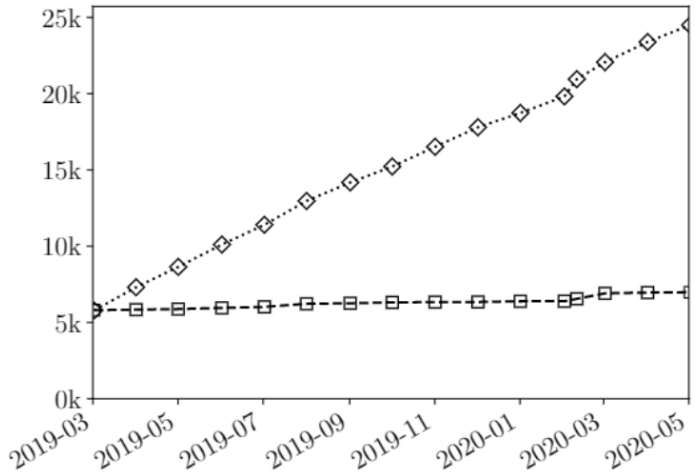
(a) BHL



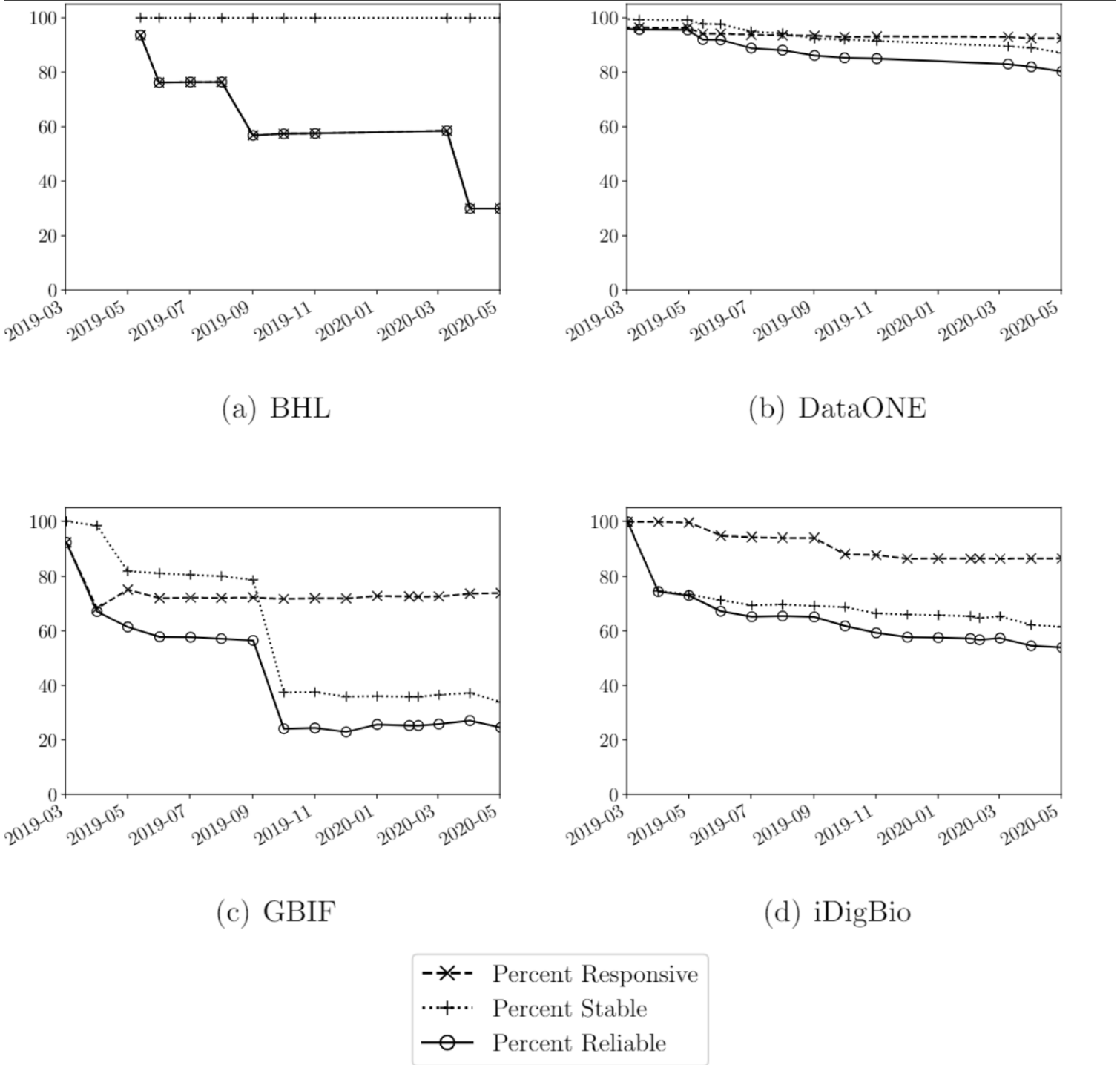
(b) DataONE



(c) GBIF



(d) iDigBio


100
80
60
40
20
0
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
2020-03
2020-05
(a) BHL
(b) DataONE
(c) GBIF
(d) iDigBio
Percent Responsive
Percent Stable
Percent Reliable

Agenda (see https://tinyurl.com/reliable-data)

1. Primer Questions (1 min)

2. Michael presents "Reliable Dataset References" (10 mins)

3. Revisit Primer Questions (1 min)

4. Discussion (45 - 12 = 33 mins)

5. Sign-up for post-discussion hallway conversation at https://ufl.zoom.us/j/98346741161 at 4PM EST (BYO coffee and cookies!)

Primer Questions

How do ***you*** make sure to keep ***your*** data around?

How do ***you*** (reliably) reference data in publications?

How do ***we*** identify ***our*** digital data so that it can be kept for the next 50 years?

Michael presents "Reliable Dataset References”

Revisit Primer Questions

How do ***you*** make sure to keep ***your*** data around?

How do ***you*** (reliably) reference data in publications?

How do ***we*** identify ***our*** digital data so that it can be kept for the next 50 years?

Post-Discussion Hallway Conversation Signup:

Please add your name / affiliation / content info via https://tinyurl.com/reliable-data and join us at https://ufl.zoom.us/j/98346741161 at 4PM EST for a casual hallway conversation.

(edited: removed email addresses to reduce exposure)

Name	**Affiliation**	**Contact Info**
Jorrit Poelen	Ronin Institute	xxx
Michael Elliott	University of Florida	xxx
Beth Chambers	William & Mary	xxx
Gil Nelson	iDigBio	xxx
Jeff Gerbracht	Cornell Lab of Ornithology	xxx

Jose Fortes	UF ACIS/iDigBio	xxx
Randy Singer	University of Michigan	xxx
Gabriel	Kamener	xxx
Ali Krzton	Auburn University	xxx
Deborah Paul	Florida State University	xxx

References

Levatich T, Padilla F (2019). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset https://doi.org/10.15468/aomfnb accessed via GBIF.org on 2019-04-08 hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4

Elliott, M. J., Poelen, J. H., & Fortes, J. (2020, January 3). Toward Reliable Biodiversity Dataset References. https://doi.org/10.32942/osf.io/mysfp . For related GBIF-mediated forum discussion: https://discourse.gbif.org/t/toward-reliable-biodiversity-dataset-references/1637/3 .

Trask B. 2015. Principles of content addressing. https://bentrask.com/?q=hash://sha256/98493caa8b37eaa26343bbf73f232597a3ccda20498563327a4c3713821df892. Accessed: 2019-12-04

NSF OAC 1839201 "EAGER: Towards the Web of Biodiversity Knowledge: Understanding Data Connectedness to Improve Identifier Practices" from the **National Science Foundation**.