Reliable Biodiversity Dataset References

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

Do current dataset referencing practices always allow the original data to be retrieved?

- Current practice (following GBIF's citation guidelines):
  - https://www.gbif.org/citation-guidelines
  - Generated by GBIF

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.

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

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

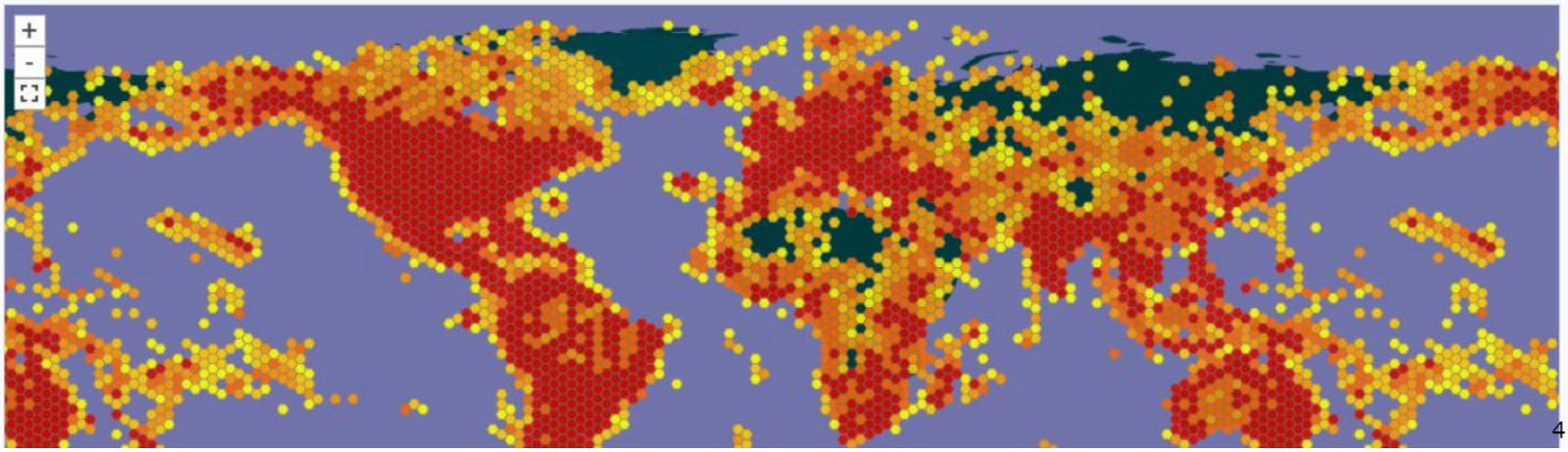
GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a [illegible]oach to citizen science by developing cooperative partnerships among ex[illegible]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

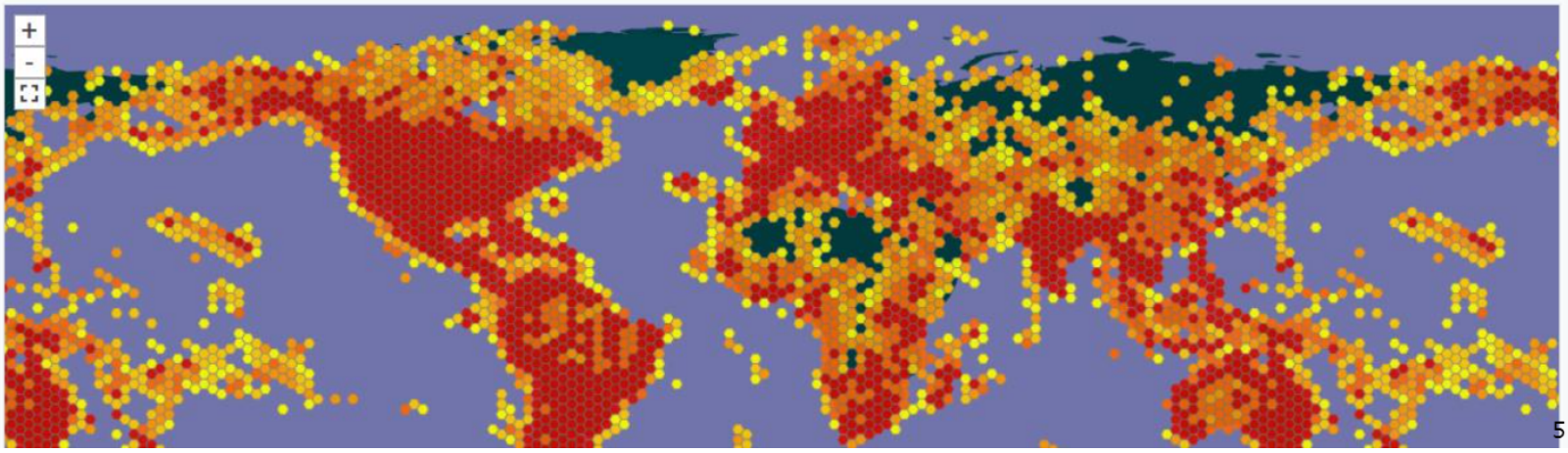
GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a[n] [...]oach to citizen science by developing cooperative partnerships among e[x] [...]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

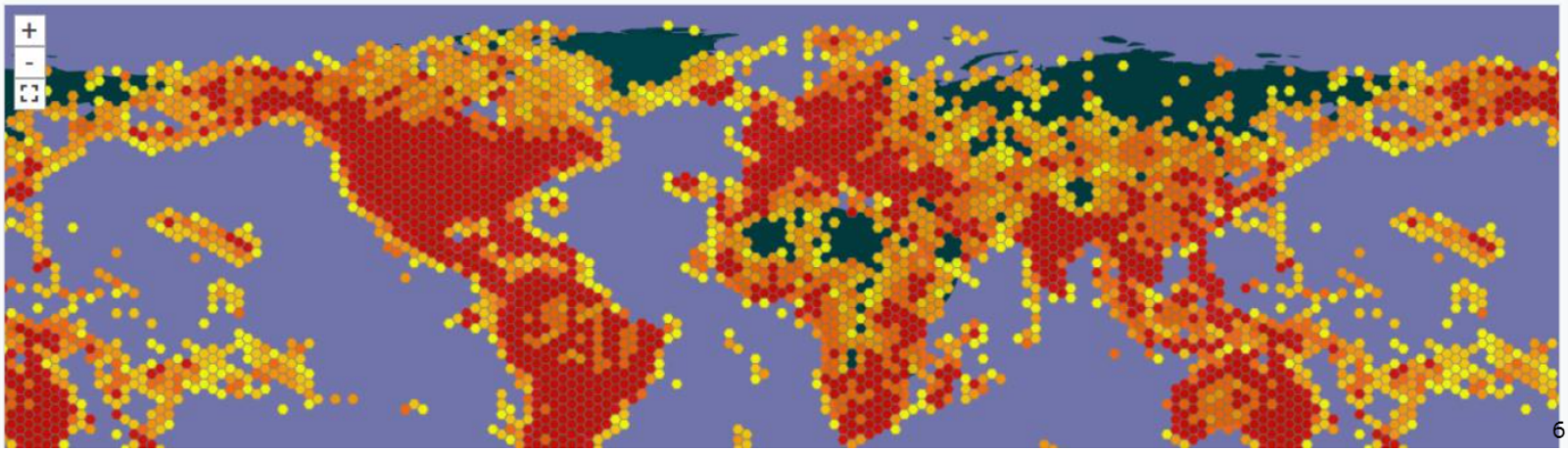
GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a ...pach to citizen science by developing cooperative partnerships among e... ...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

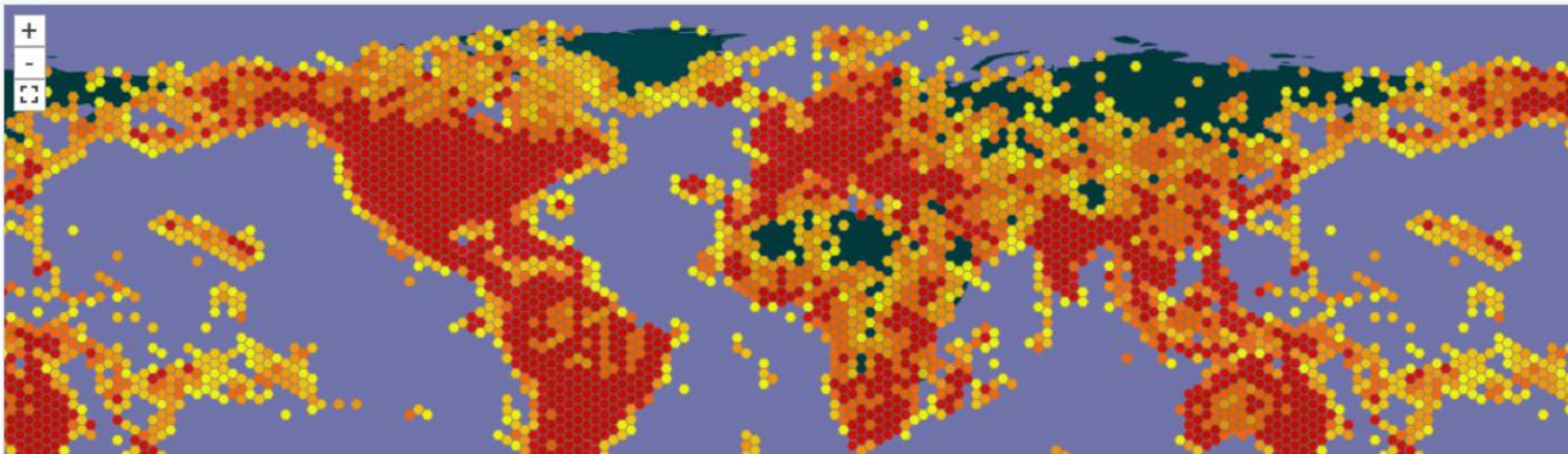
eBird is a ... itizen science by developing cooperative partnerships among e... cologis... conservation biologists, quantitative ecologists, statistician... ...s speci...ts, application developers, and data administrators. Managed by the Cornell Lab of Ornithology eBird's goal is to ...ease data quantity through participant recruitment and engagement globally, but also... More

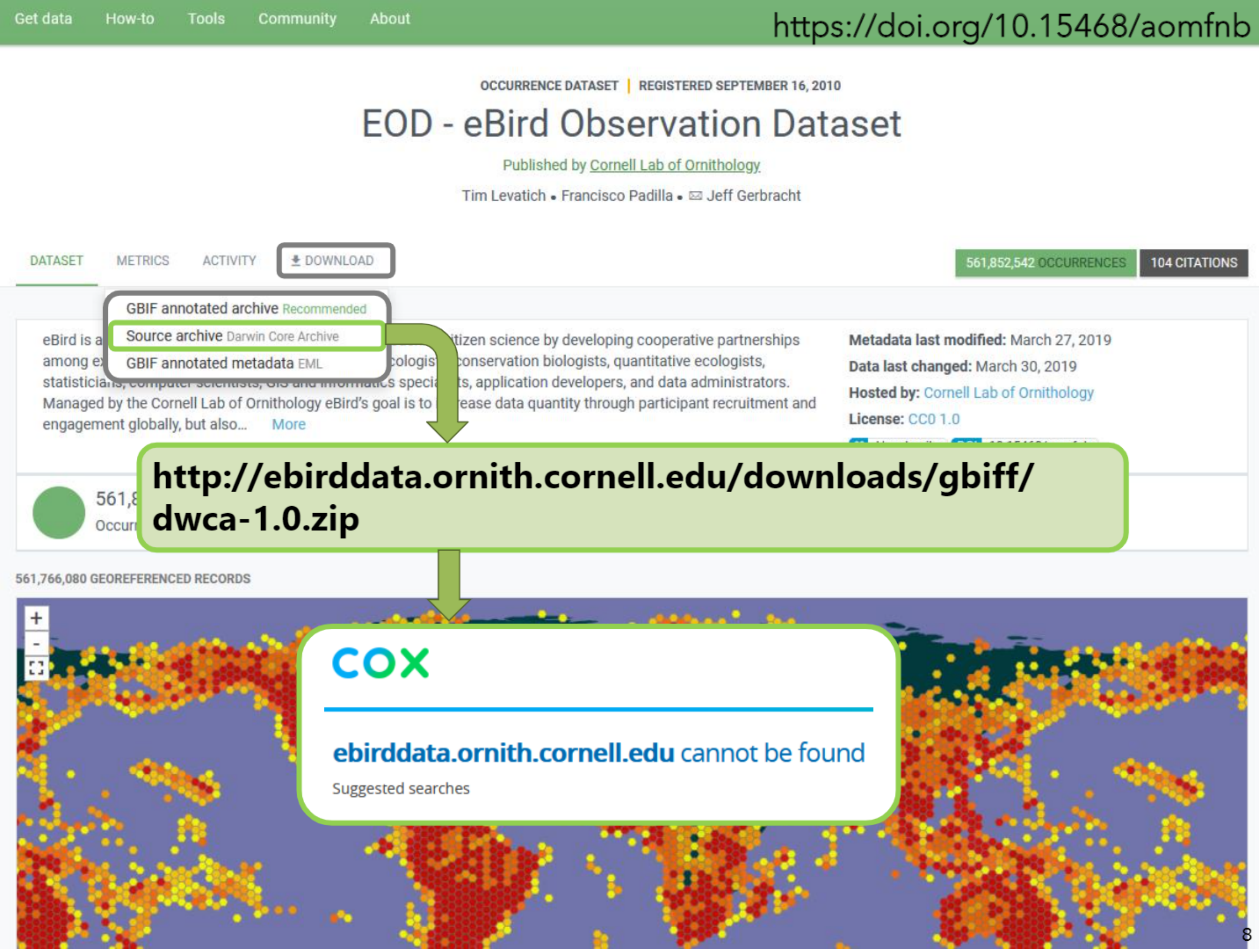
Metadata last modified: March 27, 2019
Data last changed: March 30, 2019
Hosted by: Cornell Lab of Ornithology
License: CC0 1.0

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

561,8... Occurr...

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
Managed by the Cornell Lab of Ornithology eBird's goal is to
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
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

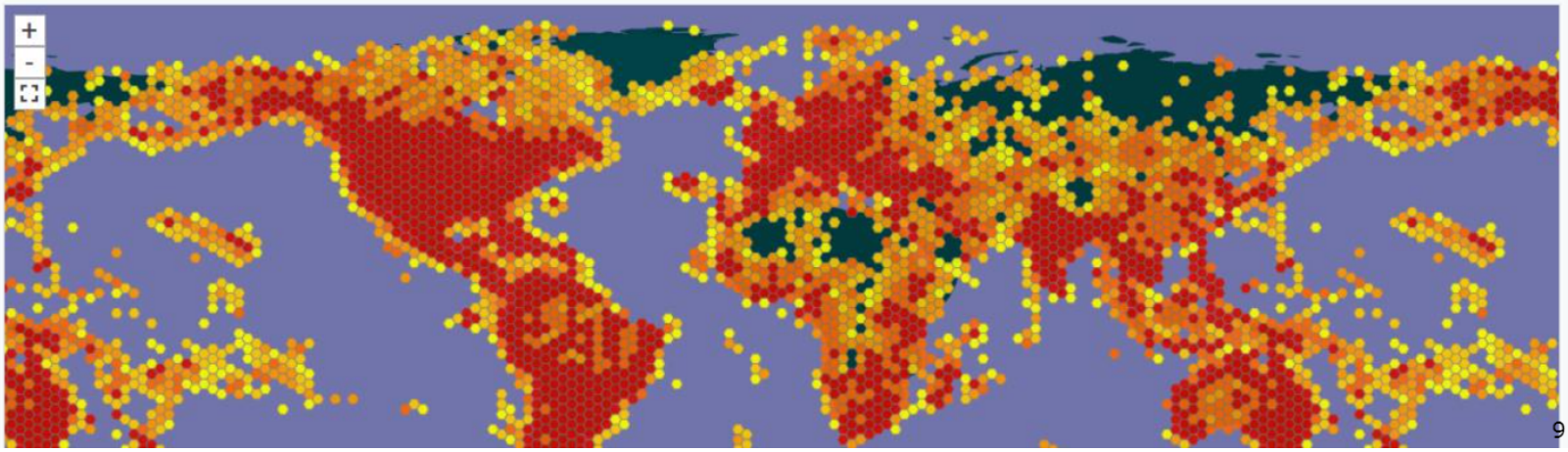
eBird is a ...pach to ...zen science by developing cooperative partnerships among e... ...cologist... ...conservation biologists, quantitative ecologists, statisticians, computer scientists, GIS and informatics specia...s, application developers, and data administrators. Managed by the Cornell Lab of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

https://www.gbif.org/occurrence/download?dataset_key=4fa7b334-ce0d-4e88-aaae-2e0c138d049e

561,766,080 GEOREFERENCED RECORDS



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…izen science by developing cooperative partnerships among e…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 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…
Occurr…

https://www.gbif.org/occurrence/download?dataset_key=4fa7b334-ce0d-4e88-aaae-2e0c138d049e

561,766,080 GEOREFERENCED RECORDS

Dataset

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…izen science by developing cooperative partnerships among e…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 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…
Occurr…

https://www.gbif.org/occurrence/download?dataset_key=4fa7b334-ce0d-4e88-aaae-2e0c138d049e

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 ...pach to ...izen science by developing cooperative partnerships among e... ...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 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...
Occurr...

https://www.gbif.org/occurrence/download?dataset_key=4fa7b334-ce0d-4e88-aaae-2e0c138d049e

561,766,080 GEORESIDENCED 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

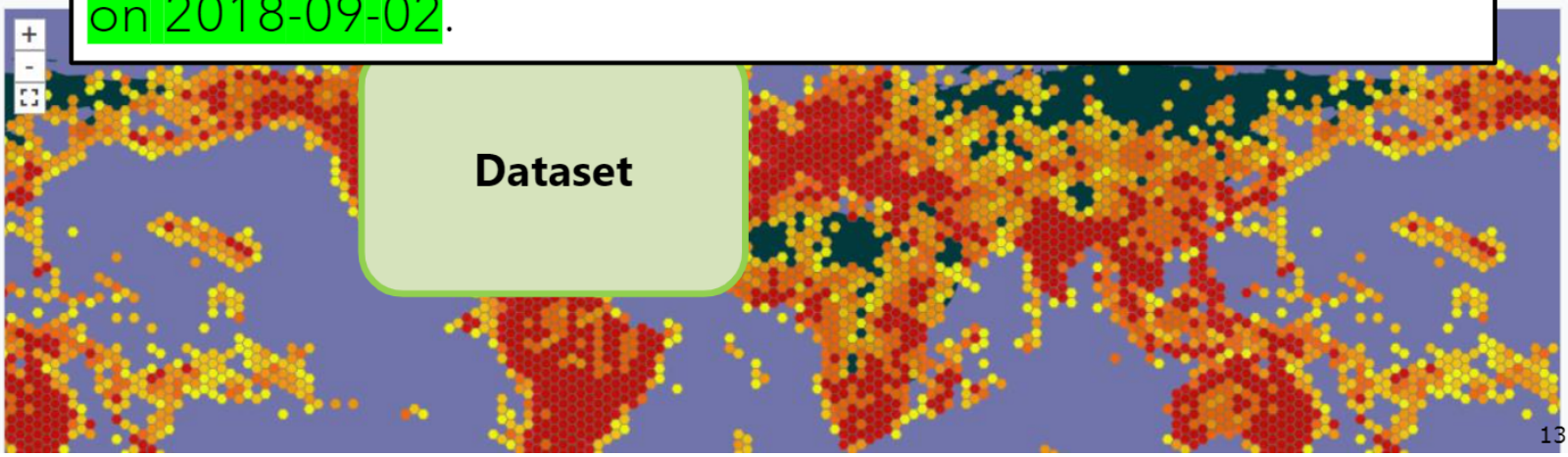
GBIF annotated archive Recommended
Source archive Darwin Core Archive
GBIF annotated metadata EML

eBird is a ...pach to ...zen science by developing cooperative partnerships among e... ...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

- In current practice, future readers must expect referenced online content to remain accessible and unaltered via a single web location (e.g. URL)

- Two potential problems:
  - **Link rot**: a link does not respond
  - **Content drift**: the content served at a link has changed
  - A "link" may be a URL, DOI, etc.
    - Note: DOIs point to URLs

- Klein et al. 2014: **one in five** Science, Technology, and Medicine articles contained references that exhibited **either link rot or content drift**
- Vision 2010: **40%** of links to supplemental materials in the Genetics journal exhibited **link rot**

- *What about in biodiversity?*

Klein M, de Sompel HV, Sanderson R, Shankar H, Balakireva L, Zhou K,
Tobin R. 2014. Scholarly context not found: One in five articles suffers
from reference rot. PLoS ONE 9:e115253.
doi:10.1371/journal.pone.0115253.

Vision TJ. 2010. Open data and the social contract of scientific publishing.
BioScience 60:330{331.
doi:10.1525/bio.2010.60.5.2.

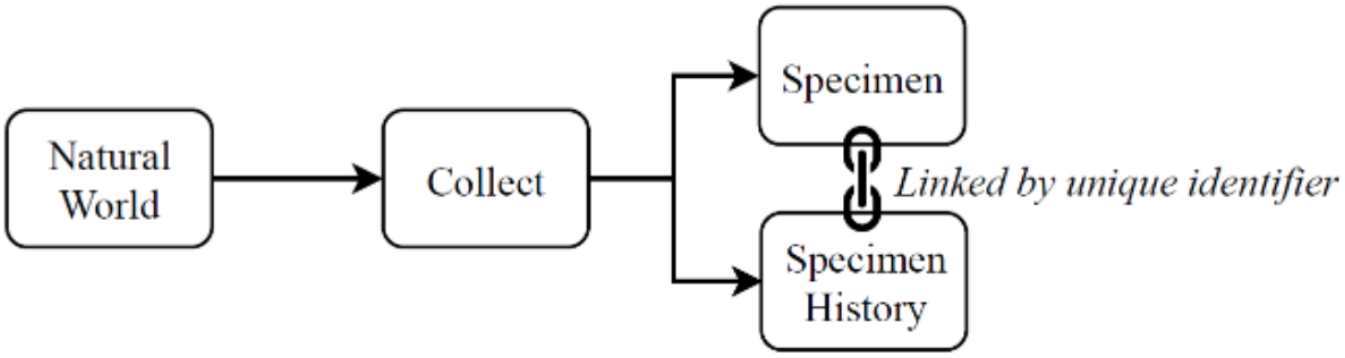
ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

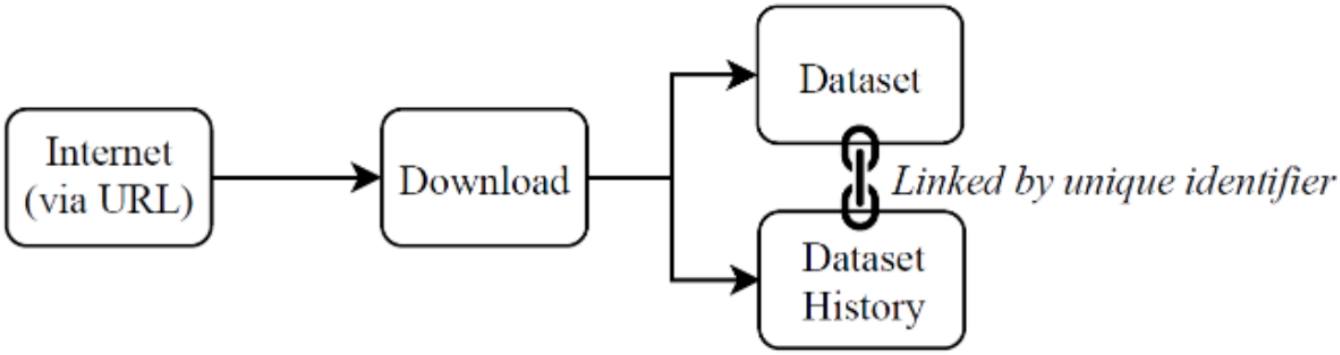
- We set out to
  - Monitor the URLs served across biodiversity data networks over an extended period of time
  - Quantify the rates of link rot and content drift in those networks

- We chose the following networks:
  - **BHL** (Biodiversity Heritage Library)
  - **DataONE** (Data Observation Network for Earth)
  - **GBIF** (Global Biodiversity Information Facility)
  - **iDigBio** (Integrated Digitized Bio Collections)

- Each data network exposes a registry of dataset URLs
  - Recall: http://**ebirddata.ornith.cornell.edu**/downloads/gbiff/dwca-1.0.zip



(a) Physical specimen collection



(b) Digital data collection

- For each dataset URL registered with biodiversity data network, we can record the following:
  - The URL
  - The date of access
  - Any content the URL provided when accessed
    - Case 1: it returned content (hopefully a dataset)
    - Case 2: the URL did not respond
- By repeating and logging such observations over time, we can build a **provenance log** for a dataset
- By collecting provenance for each dataset, we can build a provenance log for each entire data network

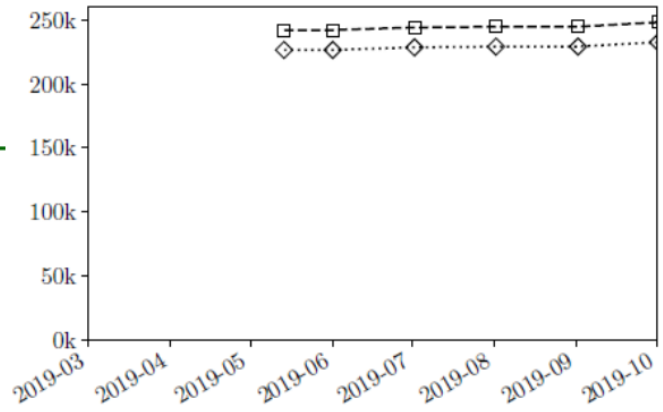
- From **March through October 2019**, we monitored all dataset URLs registered with BHL, DataONE, GBIF, and iDigBio and collected provenance logs

- Each URL was queried at the beginning of each month
  - So, each URL was queried up to eight times
  - We kept a record of the URL, the date, and the contents returned (if any) for each query

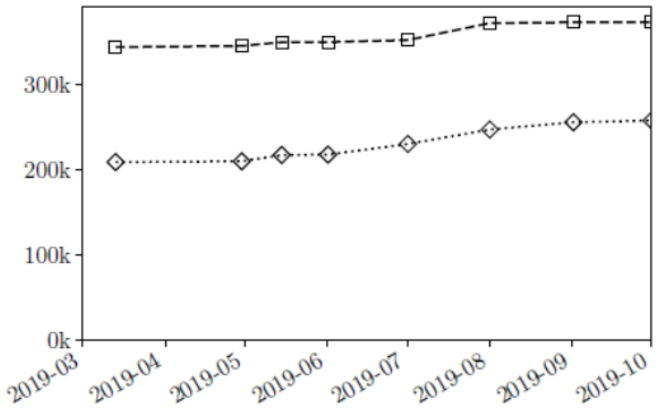
- *Note that we **only** monitored URLs that were expected to point to **datasets***

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

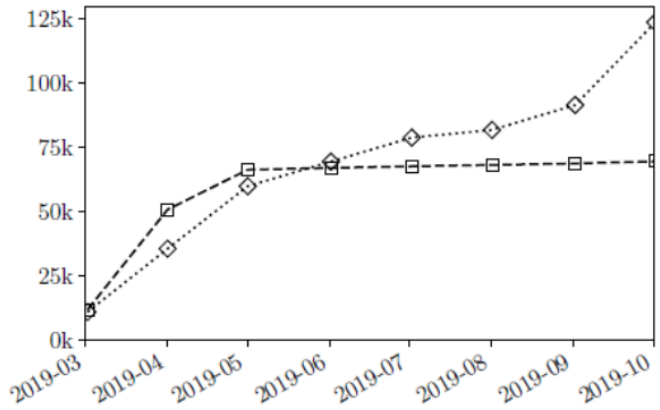
UF UNIVERSITY of FLORIDA



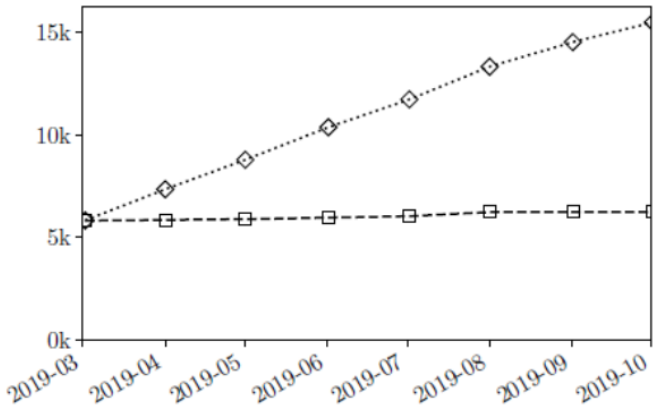
(a) BHL



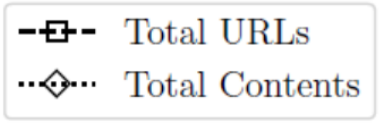
(b) DataONE



(c) GBIF



(d) iDigBio



- Some things to notice:
  - DataONE, GBIF, and iDigBio all produced new datasets (versions) at a faster pace than they produced new URLs
    - Content drift?
  - At some points in time, both GBIF and BHL exposed more dataset URLs than unique datasets
    - Link rot?
    - Multiple URLs for some datasets?

- Maybe we should take a closer look at the data
  - How often do link rot and content drift occur?
  - More generally, how reliable are the URLs?
  - How reliable is each data network?

- We define health indicators for dataset references (URLs):
  - **Unresponsive**: the link has failed to respond to one or more queries
  - **Responsive**: the link has responded to all recorded queries
  - **Unstable**: the content that the link points to sometimes changes
  - **Stable**: the content that the link points to never changes
  - **Unreliable**: the link is either unresponsive, unstable, or both
  - **Reliable**: the link is both responsive and stable

- In the provenance logs we collected, we used **hash URIs** to identify (versions of) datasets. For example:

hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c

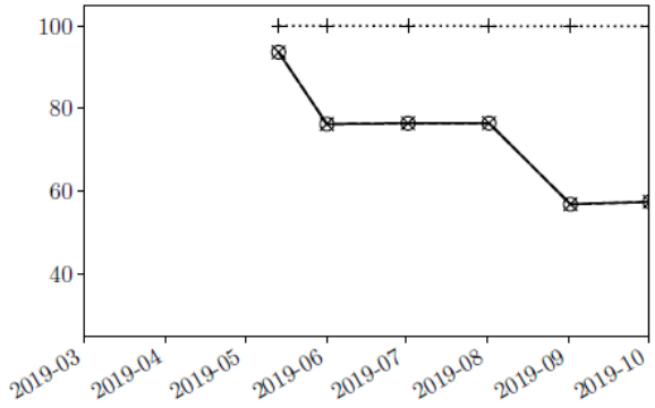
- The "hash of a dataset" is an identifier that is calculated from the data itself
- As long as the same hashing algorithm is used (e.g. SHA256), the same hash will be generated every time
  a) If the same hash is calculated from two datasets, the datasets are identical
  b) If different hashes are calculated from two datasets, the datasets are different

- When collecting provenance logs for each URL, we recorded the hash URI of the dataset returned by the URL at each point of observation
- Consider a URL: www.dataset.com/123.zip
  - February hash://sha256/aaaaaaaaaa
  - March hash://sha256/aaaaaaaaaa
  - April Did not respond
  - May hash://sha256/bbbbbbbbb
- Because the URL did not respond in April, we say the URL is **unresponsive**
- Because the hash URI changed between March and May, we say the URL is **unstable**

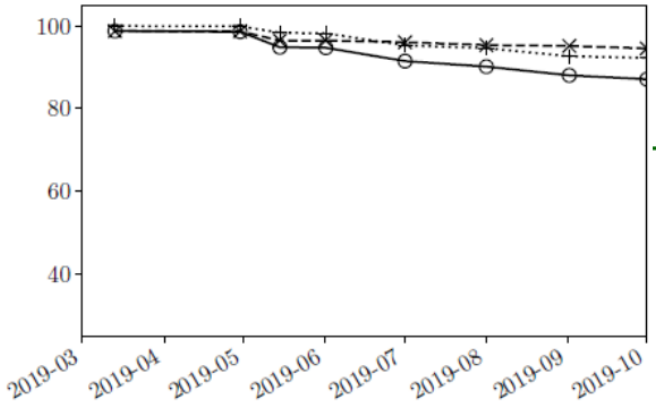
- We analyzed the provenance collected for each dataset URL within each data network
- For each data network, we calculated the
  - Percentage of (un)stable URLs
  - Percentage of (un)responsive URLs
  - Percentage of (un)reliable URLs
- Calculations were made for both monthly observations and over the full period of observation

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

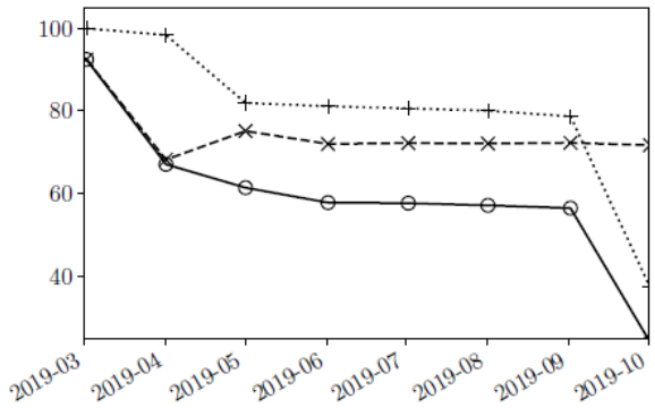
UF UNIVERSITY of FLORIDA



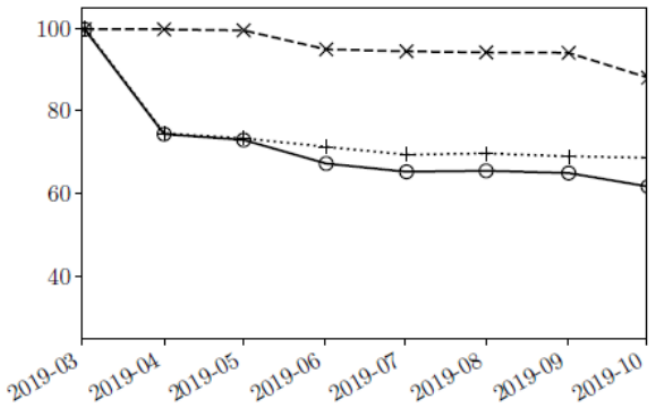
(a) BHL



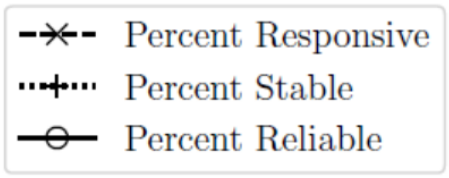
(b) DataONE



(c) GBIF

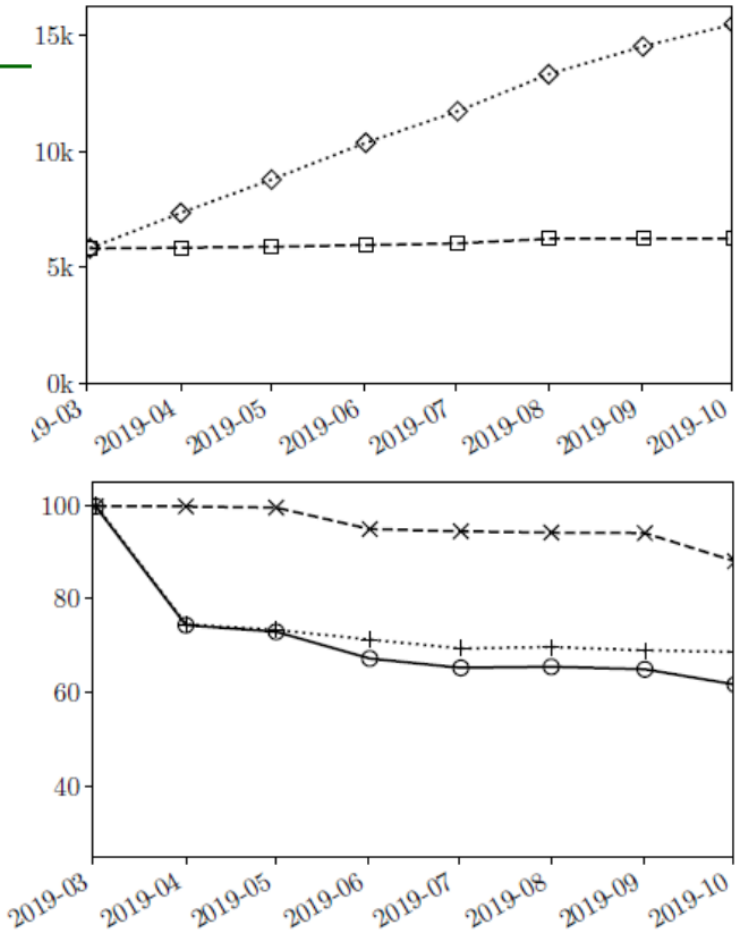


(d) iDigBio



ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA


15k
10k
5k
0k
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-09
2019-10
Total URLs
Total Contents
100
80
60
40
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-09
2019-10
Percent Responsive
Percent Stable
Percent Reliable

(d) iDigBio

Data Network	Responsive URLs	Stable URLs*	Reliable URLs
BHL[a]	57.41% (142,672)	99.97% (232,996)	57.39% (142,633)
DataONE[b]	94.55% (352,438)	92.27% (339,109)	87.09% (324,641)
GBIF[c]	71.72% (49,707)	37.35% (20,094)	24.05% (16,669)
iDigBio[c]	88.04% (5,477)	68.69% (4,251)	61.68% (3,837)
All observed URLs**	78.94% (546,645)	90.43% (593,469)	70.07% (485,203)

Table 1. Overall responsiveness, stability, and reliability for URLs observed

- Over the course of observation, across data networks:
  - 5% to 43% of URLs were at least intermittently unresponsive
  - 0% to 63% of URLs were unstable
  - 13% to 76% of URLs were unreliable

- For all data networks combined:
  - 30% of URLs were unreliable, of which
    - 48% were unstable
    - 70% were unresponsive
    - 22% were consistently unresponsive
  - For 5% of successful queries, the URL became unresponsive on the next query
  - For 4% of successful queries, the URL provided different content on the next successful query

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

- Note that the numbers we collected don't reflect URLs that were unresponsive/unstable/unreliable outside the period of observation
  - Our results are optimistic

- So, if URLs aren't a reliable referencing mechanism for biodiversity datasets, what else can we use?

- "The best way to 'future proof' an identifier scheme is to forego any intelligence within the identifier itself"
  Paskin N. 1999. Toward unique identifiers. Proceedings of the IEEE 87:1208{1227. doi:10.1109/5.771073.

- Recall the **hash URI**, e.g.

hash://sha256/29d30b566f924355a383b13cd48c3aa239d42cba0a55f4ccfc2930289b88b43c

- More generally, we can distinguish between location-based and content-based identifiers
  - URLs (and DOIs) are **location-based**
    - Locate the data for some duration
    - May identify evolving datasets
  - Hash URIs are **content-based**
    - Derived from the data
    - Always identify the same dataset and version

	Location-based identifiers	Content-based identifiers
Unique	No	Yes
Persistent	No	Yes

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 hash URI

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

- Simply adding the hash URI gives future readers the ability to verify whether the online version of the dataset is the same as the one referenced

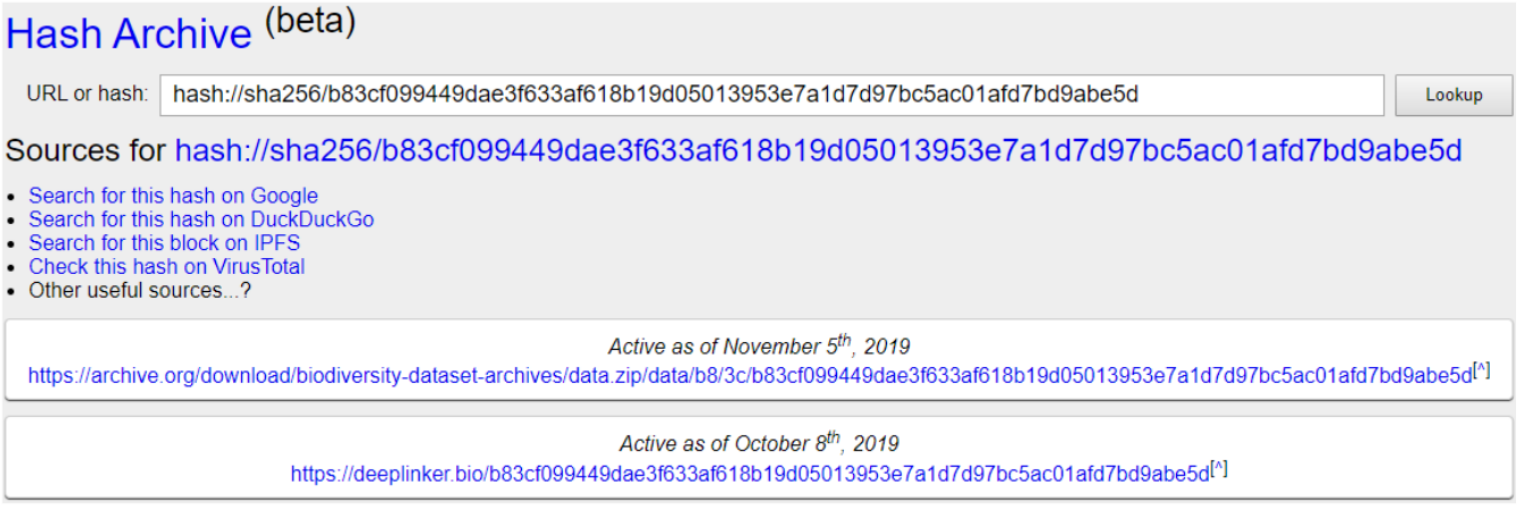
- But this still relies on URLs to find the data
  - What if the online version is different?
  - What if the referenced URL is unresponsive?

- What if we could retrieve the dataset identified by a hash URI without needing to reference a URL?

- To resolve hash URIs to datasets, we need:
  - A search index e.g. http://hash-archive.org/
    - Give it a hash URI, and it returns a known location for the identified dataset
  - A data repository e.g. Zenodo, Internet Archive, etc.
    - Provides URLs to serve datasets

- Link rot and content drift are no longer issues
  - If a URL "rots", the dataset can be uploaded to another location and registered in the search index
  - Hash URIs are specific to the exact version of the dataset they reference, so content drift is not an issue
    - The datasets retrieved from data repository can be verified by comparing with the referenced hash URI

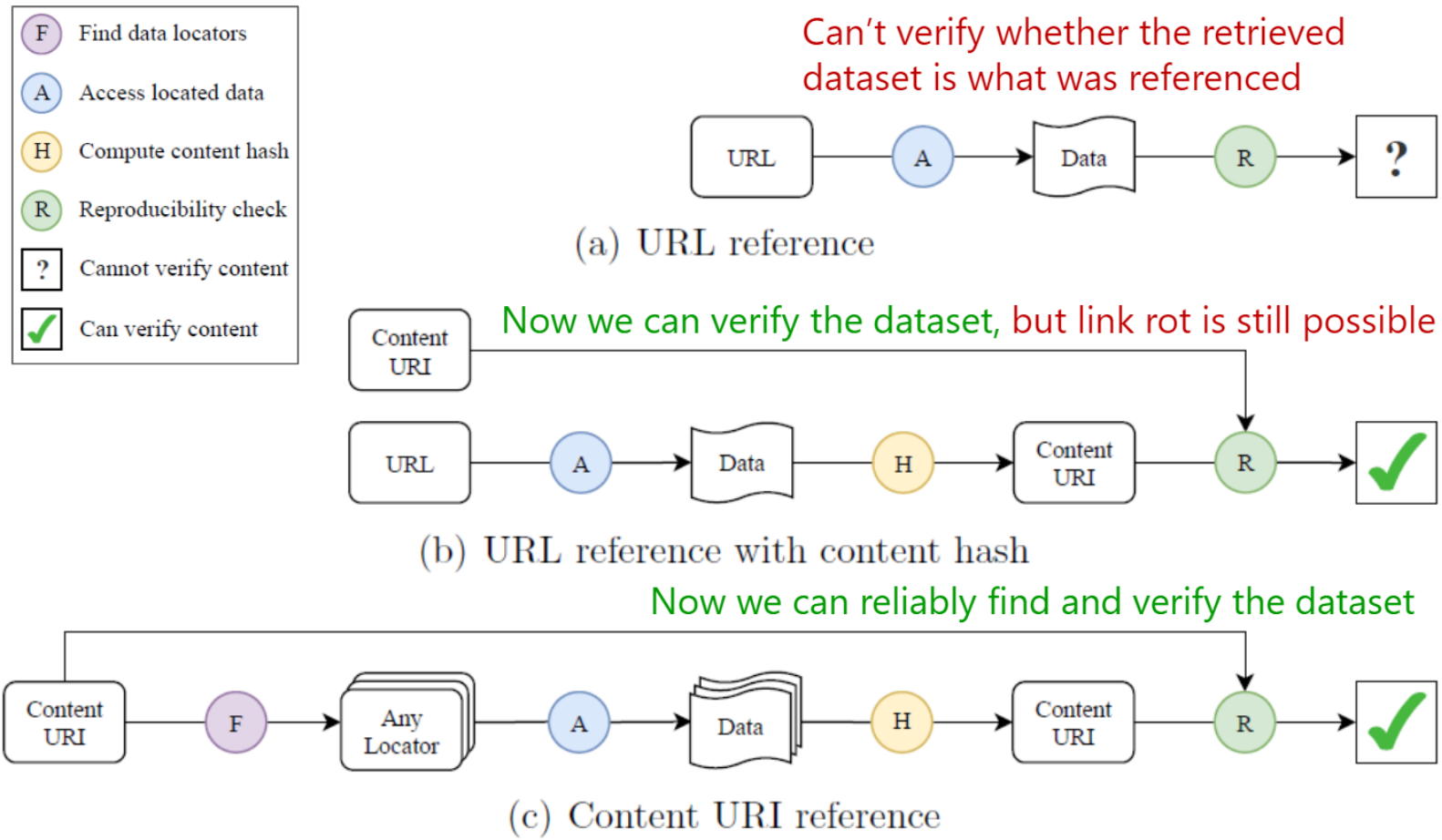
- We registered a set of hash URIs with hash-archive.org



- A query for the hash URI above returns two known locations for the referenced dataset
  - Internet Archive, archive.org
  - One of our deployed observatories

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
Now we can verify the dataset, but link rot is still possible
Content URI
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

- To preserve data:
  1) Datasets must be addressable and retrievable using content-based identifiers
  2) An agent must exist to collect datasets, record their provenance, and deposit both to data repositories
  3) Data repositories should archive data that could be used in the future
  4) Search indexes should be openly accessible and resolve hash identifiers to dataset locations within data repositories

- What about DOIs?
  - They simply point the user to URLs, which are problematic
  - However, they are still useful for identifying evolving datasets
  - Although, they rely on some DOI authority to maintain them
- What about citation files?
  - They could list, in addition to dataset URLs/DOIs, the hash URIs of the referenced datasets

- Conclusions:
  - Location-based identifiers, such as URLs and DOIs, do not reliably reference datasets
  - Content-based identifiers can reliably reference datasets
  - Data observatories can be used to collect and archive datasets from unreliable URLs
  - Archived datasets, identified by their hash URIs, can be reliably discovered distributed using decentralized search indexes and data repositories
  - All of this serves to increase the longevity of the biodiversity data record, facilitating the long-term reproducibility of scholarly works that reference evolving datasets

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