Conference

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

Title

Reliable data use in R

Type

Oral presentation, Discussion

Authors

Jorrit Poelen, Carl Boettiger

Abstract

As R is becoming a standard research tool, a basic question remains: how to reliably reference data used in R programs?

Almost any R user can sympathize with the problems of having local paths (e.g., read.csv("path/to/file.csv")), and that URLs aren't much better (or worse when bandwidth is limiting). R developers have largely sidestepped this problem by packaging the data in the code, which has made datasets like "iris" and "mcars" famous. However, this approach is of little help to any real-world data.

With help of code examples and R package "contentid", we show how to write R code that is agnostic to the location of data, works with local / remote data, and ensures that the requested data is used.

Reliable Use of Data in R

Jorrit Poelen, Ronin Institute
Carl Boettiger, UC Berkeley

1 June 2020 @ 4th Annual Digital Data in Biodiversity Research Conference
Indiana University

As R is becoming a standard research tool, a basic question remains:

How to reliably use data in R programs?

topics

> Data Use Challenges
> Toward Reliable Use of Data
> Next Steps
> Q&A

topics

> Data Use Challenges
> Toward Reliable Use of Data
> Next Steps
> Q&A

data use challenges

```
# get the data
> bird_data <- file("/home/dduck/bird_data.txt")

# count the geese
> count_geese(bird_data)
7066919
```

File paths are local **data locations** and **do not uniquely identify the content** of the data.

data use challenges

```
# get the data
> bird_data <- url("https://example.org/aves.txt")

# count the geese
> count_geese(bird_data)
7066919
```

URLs are (usually) remote **data locations** and **do not uniquely identify the content** of the data.

data use challenges

```
# get the data
> bird_data <- doi("10.5281/zenodo.3858443")
Error in doi("10.5281/zenodo.3858443") : could not find
function "doi"
```

DOIs do not resolve to **data locations** and (usually) **do not uniquely identify the data content**.

topics

> Data Use Challenges

> Toward Reliable Use of Data

> Next Steps

> Q&A

introducing content identifier

```
# get the data
> bird_data <- url("https://example.org/aves.txt")

# calculate a unique data id
> content_id (bird_data)
hash://sha256/1234...
```

Calculate a unique **content** identifier for the data using a cryptographic hash algorithm like SHA-256.

introducing content identifier

```
# get the local data
> bird_data <- file("/home/dduck/bird_data.txt")

# calculate a unique data id
> content_id (bird_data)
hash://sha256/1234...
```

Different location + same data = same content id.
Location agnostic!

introducing **content registry**

content id	**location**	**date**
hash://sha256/1234...	/home/dduck/bird_data.txt	2017-05-27
hash://sha256/1234...	https://example.org/aves.txt	2018-05-27
hash://sha256/**7765**...	/home/felix/cats.txt	2020-01-02

Now, using content ids, **create a content registry** to help find location at which data was kept at some date.

How to reliably count geese in R?

Canada Goose (Branta canadensis) via https://inaturalist.org/observations/847869
Photo 1073497, (c) Donna Pomeroy, some rights reserved (CC BY-NC)

towards Reliable Use of Data in R

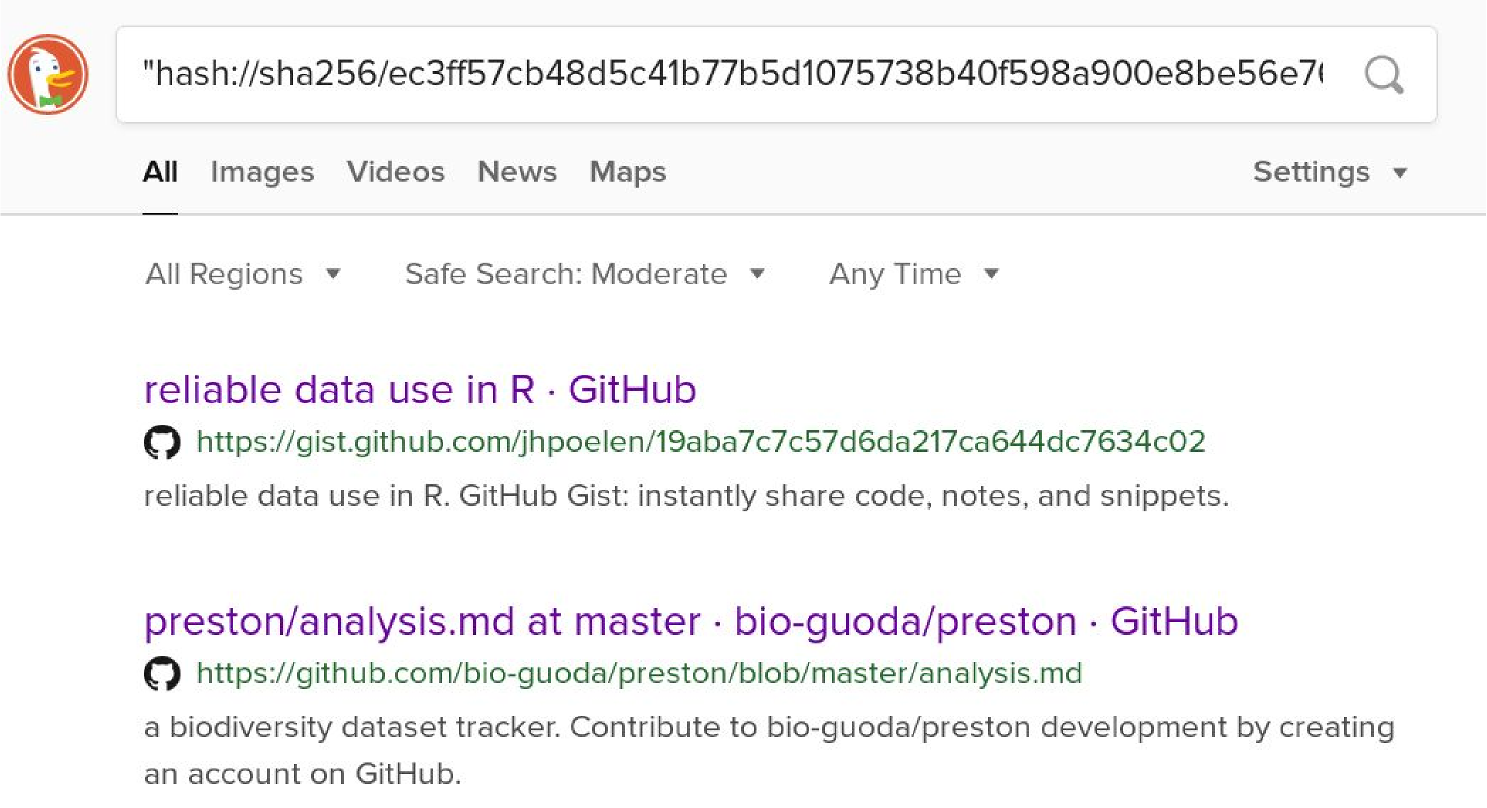
```
# step 1. register eBird* data

> contentid::register("https://zenodo.or...")
hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598
a900e8be56e7645e5a24013dffc4
```

* 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

Warning: real eBird* dataset with > 500M records

* 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

DuckDuckGo knows about it



"hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7

reliable data use in R · GitHub

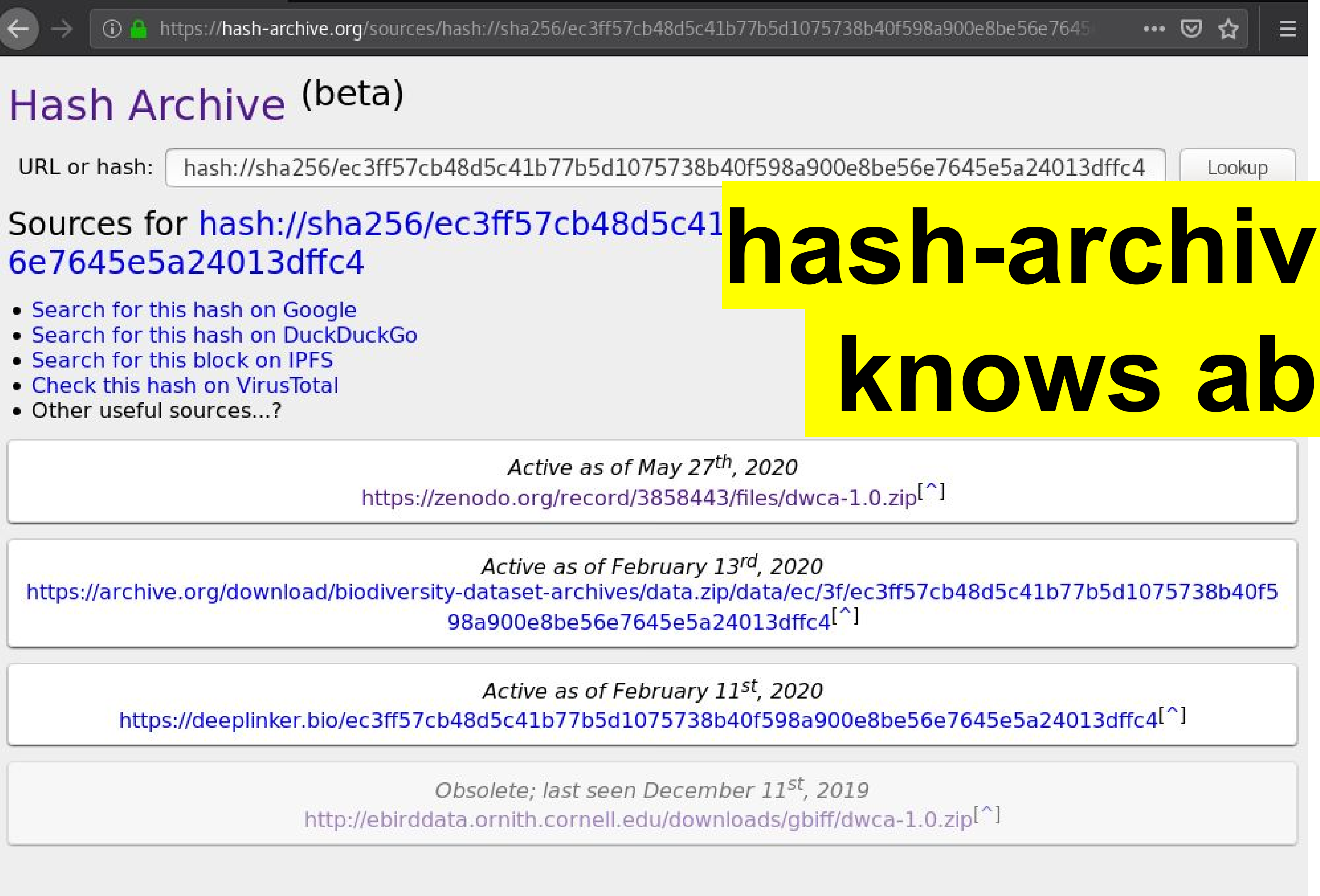
https://gist.github.com/jhpoelen/19aba7c7c57d6da217ca644dc7634c02

reliable data use in R. GitHub Gist: instantly share code, notes, and snippets.

preston/analysis.md at master · bio-guoda/preston · GitHub

https://github.com/bio-guoda/preston/blob/master/analysis.md

a biodiversity dataset tracker. Contribute to bio-guoda/preston development by creating an account on GitHub.


https://hash-archive.org/sources/hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645
Lookup

Hash Archive (beta)

URL or hash: hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4

Sources for hash://sha256/ec3ff57cb48d5c41
6e7645e5a24013dffc4

- 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...?

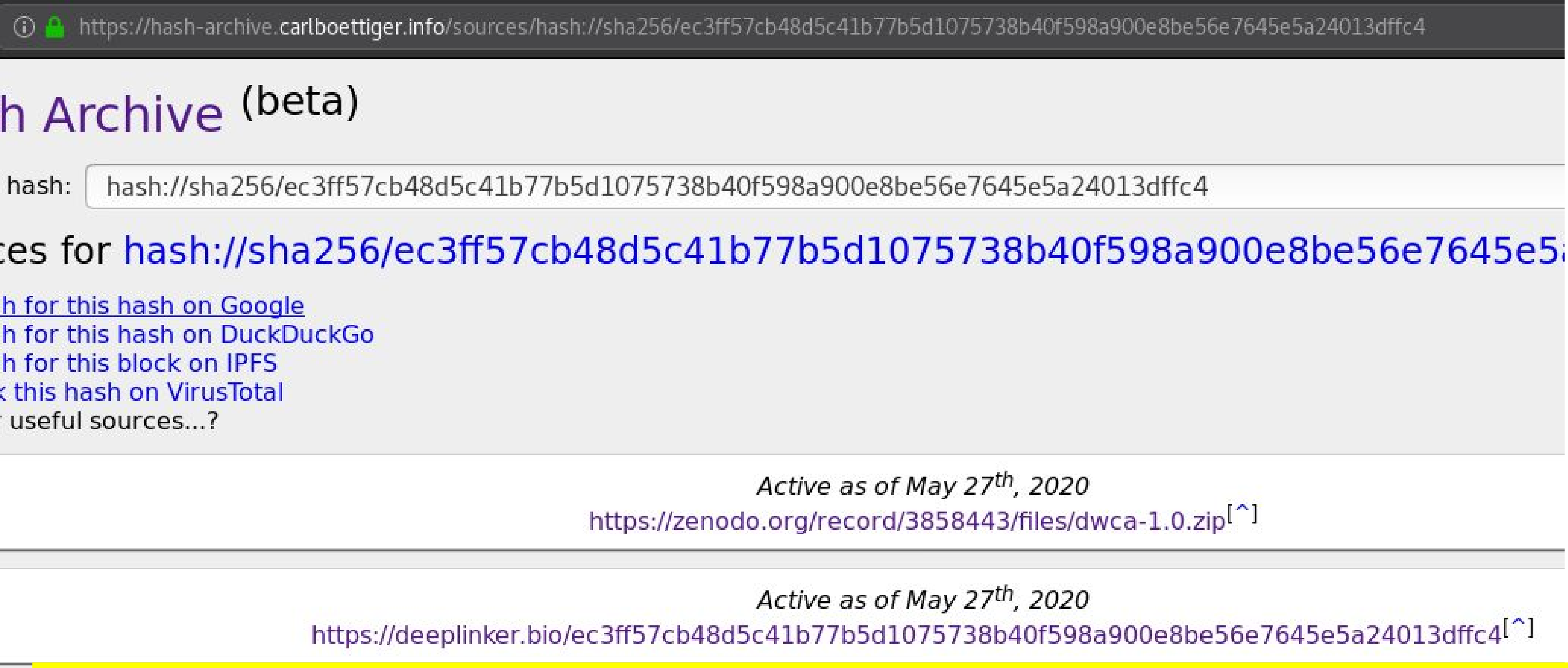
Active as of May 27th, 2020
https://zenodo.org/record/3858443/files/dwca-1.0.zip[^]

Active as of February 13rd, 2020
https://archive.org/download/biodiversity-dataset-archives/data.zip/data/ec/3f/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4[^]

Active as of February 11st, 2020
https://deeplinker.bio/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4[^]

Obsolete; last seen December 11st, 2019
http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip[^]

hash-archive.org knows about it



https://hash-archive.carlboettiger.info/sources/hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4

h Archive (beta)

hash: hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4

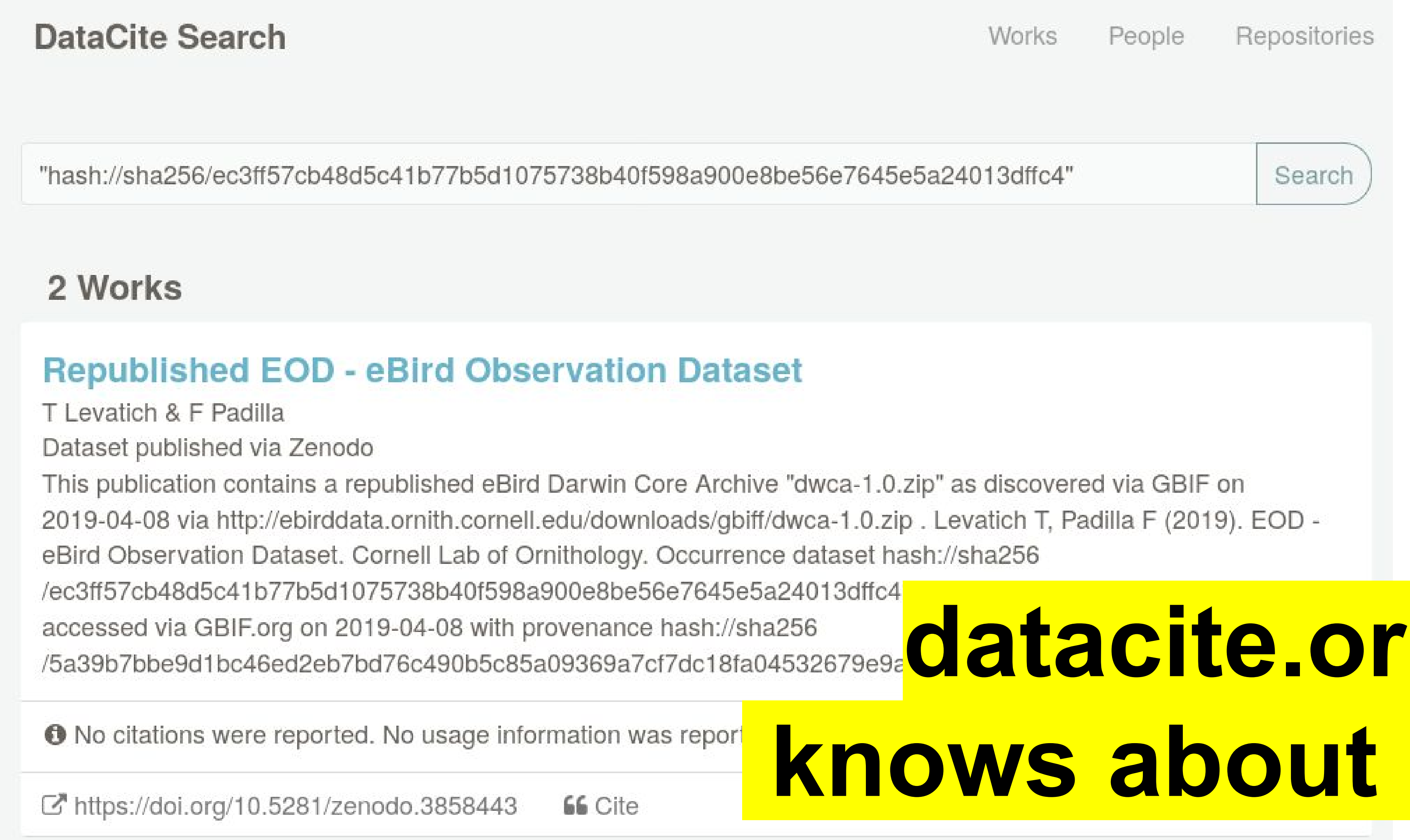
ces for hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a

h for this hash on Google
h for this hash on DuckDuckGo
h for this block on IPFS
k this hash on VirusTotal
r useful sources...?

Active as of May 27th, 2020
https://zenodo.org/record/3858443/files/dwca-1.0.zip[^]

Active as of May 27th, 2020
https://deeplinker.bio/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4[^]

hash-archive.carlboettiger.info knows about it



"hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4"

2 Works

Republished EOD - eBird Observation Dataset

T Levatich & F Padilla

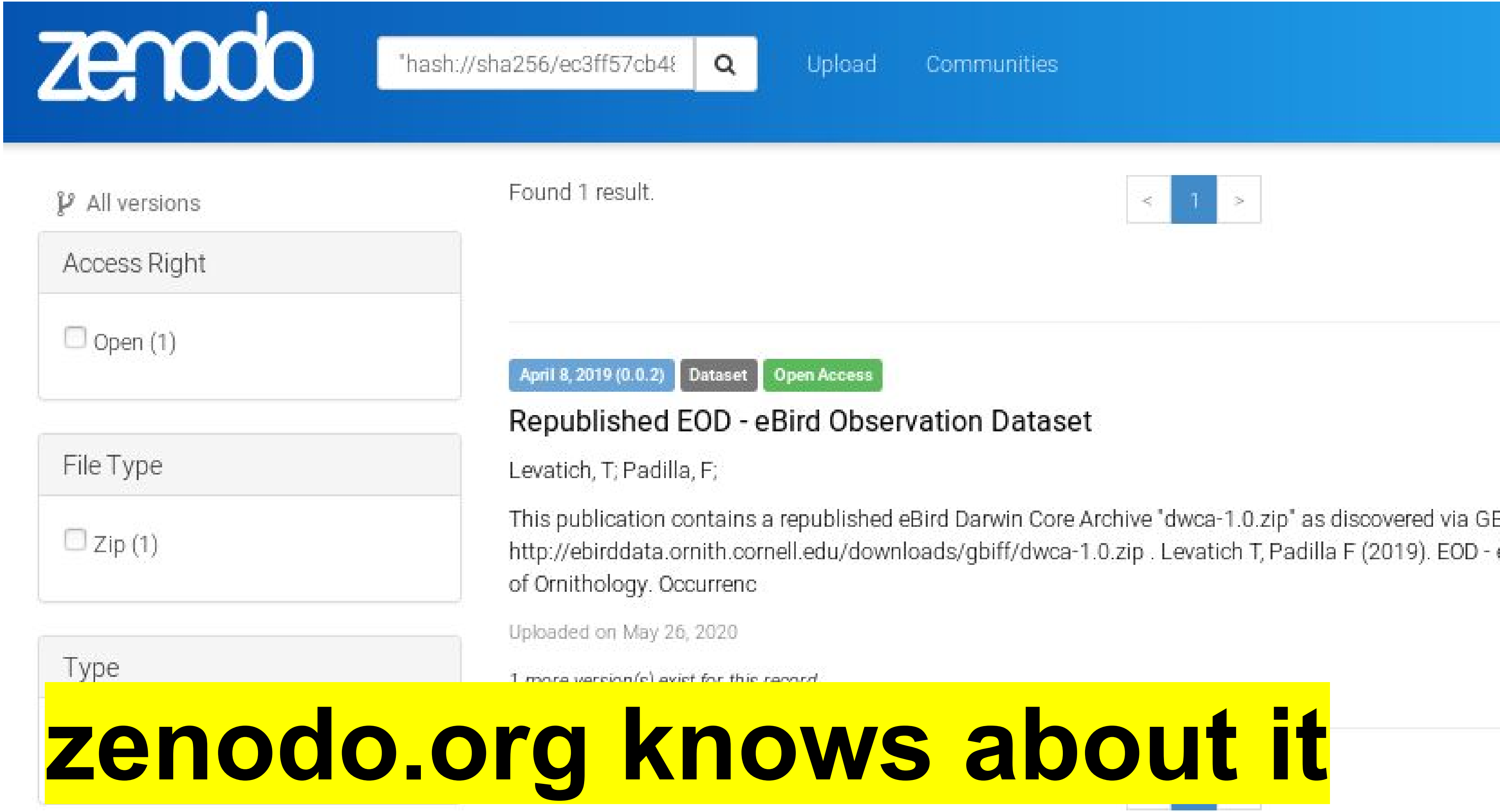
Dataset published via Zenodo

This publication contains a republished eBird Darwin Core Archive "dwca-1.0.zip" as discovered via GBIF on 2019-04-08 via http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip . Levatich T, Padilla F (2019). EOD - eBird Observation Dataset. Cornell Lab of Ornithology. Occurrence dataset hash://sha256 /ec3ff57cb48d5c41b77b5d1075738b40f598a900e8be56e7645e5a24013dffc4
accessed via GBIF.org on 2019-04-08 with provenance hash://sha256 /5a39b7bbe9d1bc46ed2eb7bd76c490b5c85a09369a7cf7dc18fa04532679e9a

ℹ No citations were reported. No usage information was report

https://doi.org/10.5281/zenodo.3858443 Cite

datacite.org knows about it



All versions

Access Right

□ Open (1)

File Type

□ Zip (1)

Type

Found 1 result.

< 1 >

April 8, 2019 (0.0.2) | Dataset | Open Access

Republished EOD - eBird Observation Dataset

Levatich, T; Padilla, F;

This publication contains a republished eBird Darwin Core Archive "dwca-1.0.zip" as discovered via GE
http://ebirddata.ornith.cornell.edu/downloads/gbiff/dwca-1.0.zip . Levatich T, Padilla F (2019). EOD - e
of Ornithology. Occurrenc

Uploaded on May 26, 2020

1 more version(s) exist for this record

zenodo.org knows about it

towards **Reliable Use of Data in R**

```
# step 1. register eBird* data

> contentid::register("https://zenodo.or...")
hash://sha256/ec3ff57cb48d5c41b77b5d1075738b40f598
a900e8be56e7645e5a24013dffc4
```

* 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

towards Reliable Use of Data in R

```
# step 2. hard code the content id in analysis script
ebird_id <- "hash://sha256/ec3ff..."

# step 3. resolve a (verified) content location
ebird_location <- contentid::resolve(ebird_id)

# step 4. count all Branta canadensis aka Canadian geese
number_of_geese <- count_geese(ebird_location)

# step 5. after successful count, add validation to script
if (number_of_geese != 7066919) { error("cannot reproduce") }
```

Current Status Reliable Use of Data in R

> **Re-used https://hash-archive.org** to (re-)register over 500GB existing biodiversity datasets across many networks and archives in in period 2018 - present
> Introduced "nouns" (e.g., **content id**) and "verbs" (e.g., **register**, **resolve**) to work towards reliable use of data in R
> Created **"contentid" R package prototype** to facilitate offline-enabled reliable data workflows

topics

> Data Use Challenges
> Toward Reliable Use of Data
> Next Steps
> Q&A

Next Steps Reliable Use of Data in R

> Adopt reliable use of data in R for own research
> Expand collaboration on content identifiers and data provenance
> Submit R package "contentid" to CRAN

Links and 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

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

Acknowledgements / Funding

an incomplete list in no particular order

Michael Elliott, Jose Fortes, Matt Collins, Jeroen Ooms and Ben Trask.

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

Canada Goose (Branta canadensis) via https://www.inaturalist.org/observations/2842153
Photo 3239787, (c) NHMLA Community Science Program, some rights reserved (CC BY-NC)

Thank you!