Signed Biodiversity Data Packages

A Method to Cite, Verify, Mobilize, and Future Proof, Large Image Corpora.

Digital Data in Biodiversity Data Conference 2023 @ Arizona State University 5-7 June 2023. doi:10.5281/zenodo.7990927

Jorrit H. Poelen https://jhpoelen.nl
Ronin Institute, UC Santa Barbara Cheadle Center for Biodiversity and Ecological Restoration
https://orcid.org/0000-0003-3138-4118

Jason Best https://fwbg.org/about-us/staff/jason-best/
Botanical Research Institute Texas
https://orcid.org/0000-0002-7414-5523

At Digital Data in Biodiversity Conference 2022, Jason Best accepted my offer to make a copy of their 13 TB image corpus for a ☕ and a 🍪.

On 05/23/22 14:43, Jorrit Poelen wrote:

> Hey Jason -
>
> Thanks for the chat earlier at the Digital Data conference.
>
> I am serious about taking on this 13 TB image corpus challenge.
>
> [...]
>
> thx,
>
> -jorrit

On 5/23/22 16:32, Jason Best wrote:

> Hello Jorrit,
>
> Thank you for the offer to provide additional back up for our images! BRIT is home to three primary and distinct plant collections which we acquired when they were orphaned by their original universities.
>
> [...]
>
> Let me know how you like your coffee and cookies.
>
> Thanks!
> Jason

On 5/23/22 16:32, Jason Best wrote:

> Hello Jorrit,
>
> Thank you for the offer to provide additional back up for our images! BRIT is home to three primary and distinct plant collections which we acquired when they were orphaned by their original universities.
>
> [...]
>
> Let me know how you like your coffee and cookies.
>
> Thanks!
> Jason

Incentivized Goal

Create 💾 of [...] three primary and distinct plant collections [...] to get ☕ and 🍪.

1. How do I efficiently access, and verify, hundreds of thousands of images?

2. How do I cite a version of a large image corpus?

1. How do I efficiently access, **and verify**, hundreds of thousands of images?

2. How do I cite a version of a large image corpus?

Botanical Research Institute of Texas (BRIT) makes their digital collections available via Darwin Core Archive URLs.

Botanical Research Institute of Texas (BRIT) makes their digital collections available via Darwin Core Archive URLs.

And, URLs are prone to link rot and content drift*.

* Elliott et al. 2020 doi:10.1016/j.ecoinf.2020.101132

Version Tracking Workflow*

step 1 of 3: Capture the BRIT DwC-A URLs.

```
echo\
 https://sernecportal.org/portal/content/dwca/VDB_DwC-A.zip\
 https://sernecportal.org/portal/content/dwca/NLU_DwC-A.zip\
 https://portal.torcherbaria.org/portal/content/dwca/BRIT_DwC-A.zip\
 | tr ' ' '\n'\
 | preston track\
 | grep hasVersion\
 | preston cat\
 [...]
```

Version Tracking Workflow*

step 2 of 3: Version DwC-As and their images

```
[...]
 | xargs preston track\
 | preston dwc-stream\
 | ./list-image-urls.sh\
 | xargs -L25 preston track
```

* see https://github.com/bio-guoda/preston-brit-2022 for more info

Version Tracking Workflow*

step 3 of 3: Wait for workflow to complete.

Version Tracking Workflow*

step 3 of 3: Wait for workflow to complete.

.

Two Weeks

Version Tracking Workflow*

step 3 of 3: Wait for workflow to complete.

. . .

Four Weeks

Version Tracking Workflow*

step 3 of 3: Wait for workflow to complete.

. . .

Six Weeks

Version Tracking Workflow*

step 3 of 3: Wait **two months** for workflow to complete.

. . . .

Eight Weeks

(Verifiable) BRIT Stats @ 2022-06-06

```
preston history\
 --remote https://linker.bio\
 --anchor hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d81e233299baf5e94\
 | tail -n1 | preston cat | grep zip\
 | preston dwc-stream --remote https://linker.bio\
 | jq -c 'select(.["http://rs.tdwg.org/dwc/terms/basisOfRecord"] == "PreservedSpecimen")'\
 | wc -l
```

990632 (~1.0M preserved specimen records)*

```
preston history\
 --remote https://linker.bio\
 --anchor hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d81e233299baf5e94\
 | tail -n1 | preston cat | grep zip\
 | preston dwc-stream --remote https://linker.bio\
 | jq -c 'select(.["http://purl.org/dc/terms/type"] == "StillImage")'\
 | wc -l
```

826195 (~0.8M still image records)*

* as of 2023-05-30, same URLs yielded 1155419 (~1.2M) preserved specimen records and 923363 (~0.9M) still image records

1. How do I efficiently access, and verify, hundreds of thousands of images?

2. How do I cite a version of a large image corpus?

Cite BRIT Image Corpus

```
preston qrcode\
 --remote https://linker.bio\
 --anchor hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d81e233299baf5e94\
 > qrcode.png
```

Elliott et al. 2023 *Sci Data* doi:10.1038/s41597-023-02230-y

Cite Individual Image

```
preston head --anchor\
 hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d81e233299baf5e94\
 | preston cat\
 | tail -n1 | grep -oE 'hash[^>]*'
```

hash://sha256/baefd416a0122a254dd68b97e41ada80764f5a0fcb13d626429b3e08403a4bb2

Cite Individual Image Metadata

```
preston history\
 --remote https://linker.bio\
 --anchor hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d81e233299baf5e94\
  | tail -n1 | preston cat | grep zip\
  | preston dwc-stream --remote https://linker.bio\
  | grep "00-bGGY6Kb3fK4hTT6CPQSp5P/resize:1250/format:jpeg"\
  | jq --raw-output '.["http://www.w3.org/ns/prov#wasDerivedFrom"]'
```

line:zip:hash://sha256/371984ca4566b7b6bc760d0766873b469e12af2d87ce9218f1da888a1b4c3948!/multimedia.csv!/L134160

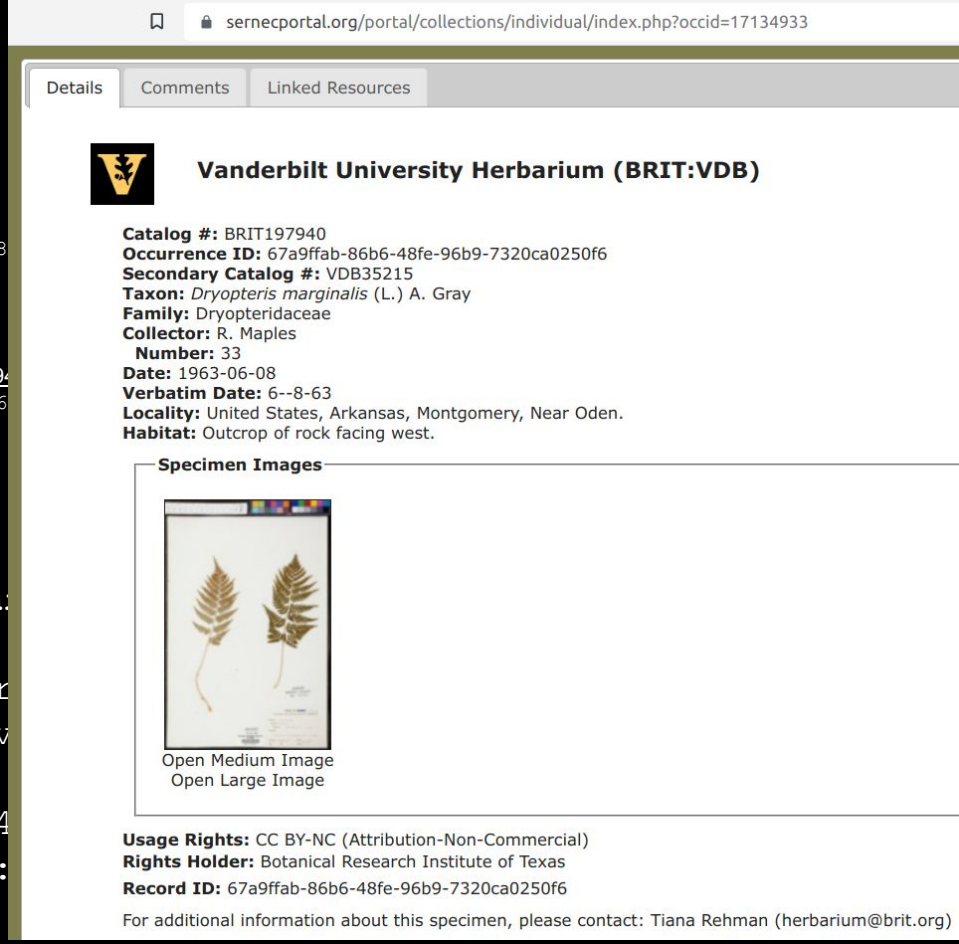
```
curl 'https://linker.bio/line:zip:hash://sha256/371984ca4566b7b6bc760d0766873b469e12af2d87ce9218f1da888a1b4c3948!/multimedia.csv!/L1,L134160'\
 | mlr --icsv --oxtab cat
```

coreid	**17134933**
[...]	
Owner	**Vanderbilt University Herbarium (VDB)**
[...]	
UsageTerms	CC BY-NC (Attribution-Non-Commercial)
WebStatement	http://creativecommons.org/licenses/by-nc/3.0/
[...]	
providerManagedID	urn:uuid:bd1740da-ea8c-489f-b3c4-8f0a7b87affd
MetadataDate	2018-01-02 12:35:50
format	image/jpeg
associatedSpecimenReference	**https://sernecportal[...]?occid=17134933**
type	StillImage

Cite Individual Imag

```
preston history\
 --remote https://linker.bio\
 --anchor hash://sha256/76d40abccfc71bc2cdaf4ea4a6003b9ac49123b27abe9f0d8
 | tail -n1 | preston cat | grep zip\
 | preston dwc-stream --remote https://linker.bio\
 | grep "00-bGGY6Kb3fK4hTT6CPQSp5P/resize:1250/format:jpeg"\
 | jq --raw-output '.["http://www.w3.org/ns/prov#wasDerivedFrom"]'
line:zip:hash://sha256/371984ca4566b7b6bc760d0766873b469e12af2d87ce9218f1da888a1b4c39
curl 'https://linker.bio/line:zip:hash://sha256/371984ca4566b7b6bc760d076
 | mlr --icsv --oxtab cat
```

```
coreid                          17134933
[...]
Owner                           Vanderbilt Un
[...]
UsageTerms                      CC BY-NC (Attr
WebStatement                    http://creativ
[...]
providerManagedID               urn:uuid:bd174
MetadataDate                    2018-01-02 12:
format                          image/jpeg
associatedSpecimenReference     https://sernecportal[...]?occid=17134933
type                            StillImage
```



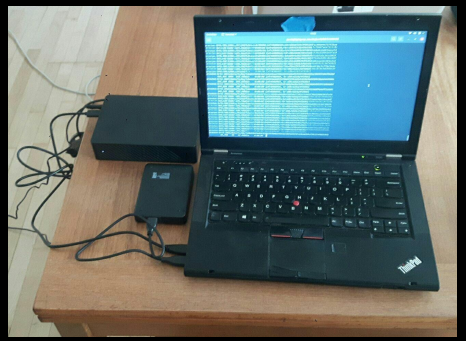
How to transfer our *verifiable* BRIT image corpus using an openly accessible reliable transportation infrastructure?

Transfer Rates

Method	Duration	Rate (image/s)	Locality
Web API	2 months	0.2	US > Germany
rsync via internet*	about a day	10	Germany > MN 55406
rsync via USB 3.0*	about an hour	250	Minnesotan kitchen table
US Postal Service*	3 days**	3	MN 55406 > TX 76107

* Signed corpus can be independently verified after transfer.
** Transfer included a weekend.

Jorrit receives storage media from Jason.



Jorrit transfers BRIT corpus onto storage media.

Jorrit labels the storage media.

Jorrit boxes storage media for shipping.


UNITED STATES POSTAL SERVICE.

POWDERHORN
3045 BLOOMINGTON AVE
MINNEAPOLIS, MN 55407-9998
(800)275-8777

09/16/2022 01:05 PM

Product	Qty	Unit Price	Price
Prepaid Mail	1		$0.00

Fort Worth, TX 76107
Weight: 3 lb 10.80 oz
Acceptance Date:
Fri 09/16/2022
Tracking #:
9405 8036 9930 0813 4150 91

USPS acknowledges receipt.

Jason copies verified corpus onto BRIT servers.

Incentivized Goal

Create 💾 of [...] three primary and distinct plant collections [...] to get ☕ and 🍪.

Incentivized Goal

Create 💾 of [...] three primary and distinct plant collections [...] to get ☕ and 🍪.

order #199255707 / 2022-10-07T08:15-05:00 @ Hark! Cafe, Minneapolis, MN

1. How do I efficiently access, and verify, hundreds of thousands of images? **Version, package, and sign.**

2. How do I cite a version of a large image corpus? **With a data package signature.**

Methods

MJ Elliott, JH Poelen, JAB Fortes (2020). Toward Reliable Biodiversity Dataset References. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2020.101132

MJ Elliott, JH Poelen, JAB Fortes (2023). Signing data citations enables data verification and citation persistence. *Scientific Data*. https://doi.org/10.1038/s41597-023-02230-y

1. How do **you** efficiently access, and verify, hundreds of thousands of images?

2. How do **you** cite a version of a large image corpus?

Funding / Acknowledgments

Jason Best, USPS, Hark! Cafe in Minneapolis, NSF OAC 1839201 and iDigBio for facilitating discussions on biodiversity informatics.