A Versioned Literature Corpus Derived from Biodiversity Heritage Library hash://md5/53e144641ffded6800dea502a8bb47ed

(a working document)

Jorrit H. Poelen Donat Agosti

2024-09-23

Abstract

The Biodiversity Heritage Library (BHL) contains over 300k items of openly accessible digital/digitized works on topics such as biodiversity, ecology and biology. While these works are openly available through the internet, extra work is needed to make the digitized versions of the scholarly works citable, verifiable and reusable. We used Preston, a biodiversity data tracker, to help track, version, package, and mobilize, a subset of BHL to produce a 190 GiB corpus containing over 30k digital representations of works along with their associated metadata and origin. With this, we compiled a corpus of scholarly works that can be independently verified and securely cited regardless of the digital communication method used to transmit or store the digital versions of the works. Citable and independently verifiable corpora like these are essential ingredients for reproducable data integration workflows as well as machine learning models and other so-called artificial intelligence algorithms.

pdf / docx / md

edit source / share suggestions

Biodiversity Heritage Library (BHL, https://biodiversitylibrary.org) contains hundreds of thousands of digital works related to biodiversity.

This project, "Bridging Biodiversity Heritage Library (BHL) to Biodiversity Literature Repository (BLR)" [1], aims to create a reusable corpus of verifiable digital representations (or digital copies) of existing biodiversity literature. This citable and redistributable collection is anchored in pdfs of known provenance (or origin). In addition, the collection associates metadata to these digital artifacts.

[1] Poelen, J. H. (2024). A Versioned Literature Corpus derived from Biodiversity Heritage Library hash://md5/53e144641ffded6800dea502a8bb47ed (0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13377084

This metadata includes author names, journal, publication year, DOI [2], urls and other identifying information.

With this, this corpus creates an explicit, verifiable association between a digital artifacts (the pdf) and their associated metadata, and allows for signed data citations [3]. These signed data citations enrich the existing associations with DOIs (or other identifiers) issued by publishers or other entities with verifiable links that do no rely on the complex sequence of dynamic redirection provided through the DOI/Handle System [4]. By decoupling the digital representations of works, we allow for verification of digital corpora independent of the type of storage media or communication method (e.g., spinning magnetic disk, http/tcp, Handle system) used.

The current version of this corpus contains a versioned subset of metadata associated to specific titles (see below), covering over 30k items with references to copies of their associated pdfs.

Included titles are:

```
Records of the Indian Museum
Revue suisse de zoologie
Bulletin of the Museum of Comparative Zoology at Harvard College
Journal of mammalogy
Proceedings of the Zoological Society of London
Archivos do Museu Nacional do Rio de Janeiro
The journal of the Bombay Natural History Society
Proceedings of the Biological Society of Washington
Bulletin of the British Museum (Natural History) Zoology
Atti della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale in Mil
Atti della Società italiana di scienze naturali e del Museo civico di storia naturale di Mil
Bonner zoologische Beiträge : Herausgeber: Zoologisches Forschungsinstitut und Museum Alexan
Zeitschrift für Säugetierkunde : im Auftrage der Deutschen Gesellschaft für Säugetierkunde e
Annals of the Carnegie Museum
Records of the Western Australian Museum
Fieldiana. Zoology
Occasional papers of the Museum of Natural History, the University of Kansas
Zoologischer Anzeiger
Proceedings of the United States National Museum
```

[2]Digital Object Identifiers (DOI) are digital identifiers for any kind of object, physical, digital or abstract. Through the "Handle System", these DOIs may be used on the Internet as a way to query for available metadata on the referenced object via a complex redirection scheme using DNS, TCP/IP, and a centralized service operating on https://doi.org .

[3]Elliott M.J., Poelen, J.H. & Fortes, J.A.B. (2023) Signing data citations enables data verification and citation persistence. Sci Data. https://doi.org/10.1038/s41597-023-02230-y hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d

[4]Poelen, J. H. (2024, August 19). Bug Pictures Beyond The Internet. Zenodo. https://doi.org/10.5281/zenodo.13350983 https://jhpoelen.nl/arcadia-talk-2024-08-19

Example

An example digital artifact in this versioned literature corpus is the pdf with signature or content id hash://md5/45657d5177e716a2c339f4e6a3bb4f94 . This content is associated with:

> Wirth, W. W. (1953). Biting midges of the heleid genus Stilobezzia in North America. Proceedings of the United States National Museum, 103, 57–85. https://www.biodiversitylibrary.org/part/71960 hash://md5/45657d5177e716a2c339f4e6a3bb4f94 https://doi.org/10.5281/zenodo.13682901 https://doi.org/10.5479/si.00963801.103-3316.57].

A copy of this pdf was retrieved from https://www.biodiversitylibrary.org/partpdf/71960 on 2024-08-26. And, through metadata provided by BHL in the RIS format, we have the following RIS record associated with the digital artifact, as retrieved via https://linker.bio/line:zip:hash://md5/4d71f93adf6d6b1ec1f06bf726318029!/data/bhlpart.ris!/L832925-L832938 :

```
TY  - JOUR
TI  - Biting midges of the heleid genus Stilobezzia in North America
T2  - Proceedings of the United States National Museum
VL  - 103
IS  - 3316
UR  - https://www.biodiversitylibrary.org/part/71960
PB  - Smithsonian Institution Press, [etc.]
CY  - Washington
PY  - 1953
SP  - 57
EP  - 85
DO  - 10.5479/si.00963801.103-3316.57
AU  - Wirth, Willis Wagner
ER  -
```

A copy of this pdf may be retrieved via Zenodo, or any other system that allows for looking up content by their digital signature (e.g., hash://md5/45657d5177e716a2c339f4e6a3bb4f94).

For instance, via:

```
preston cat\
 --remote https://linker.bio,https://zenodo.org\
 hash://md5/45657d5177e716a2c339f4e6a3bb4f94\
 > wirth1953.pdf
```

or via an alternate point-and-click query by checksum (aka content hash, content id):

```
https://zenodo.org/search?q=_files.checksum%3A%22md5:45657d5177e716a2c339f4e6a3bb4f94%22&f=a
```

Alternatively, the internet can be searched for any other associated identifier (e.g., https://www.biodiversitylibrary.org/part/71960, https://doi.org/10.5479/si.00963801.103-3316.57. However, as these (meta) data associated with these identifiers cannot be independently verified, the content was originally used and cited may have changed or disappeared altogether because of content drift or link rot.

Technical info

Files included in this publication are:

HEAD - described the signature of this literature corpus as a MD5 hash.

titles.txt - list of a the BHL titles considered in this corpus

items.txt - list of items considered in this corpus

as well as the following files named after the content their contain using the md5 hash algorithm [5]:

```
data/a4/45/a445e6eb1138710be4b93db45ab7b4a9
data/d2/68/d268444ddf44545c53e6bf4635b3ae57
data/98/fb/98fbf06892fcf919bc2107446147cc4a
data/1f/b7/1fb70a2c683eb7a5545900bc6668950f
data/fe/b6/feb6470b553aa9e104a6d7114a424a05
data/e1/9a/e19ad69093efab2b86c3ca04de8f95ba
data/b9/01/b9014b0b403516b1e852748d72633c7d
data/72/90/7290c3e847ed1d10e47e4fd12188b752
data/63/af/63af62d86d0f3d4d26bfd534bf7282bb
data/01/6c/016c9f11f52e6ce67c7399a4d588c2eb
data/a2/8a/a28a615337c2c187a5d3286923642ad6
data/a2/50/a250d215282393ad43494dd9de11da4a
data/ac/44/ac44c9ac6746a9fe2babaac45ec58860
data/3e/43/3e437559b980d2e88daf3c05ec2f4d77
data/05/02/0502e4ea67c725e22f71aef16b164303
data/53/e1/53e144641ffded6800dea502a8bb47ed
data/b8/55/b855da1ec4dd0600c8fd736a0f413ec2
data/35/2f/352fa9f7e46576462cf7840b2992de0d
data/f2/06/f206133d0b42fc9126e36a30654748ec
data/d8/6a/d86a121f36f8e6f081a20803444bc1aa
```

Technical info

This part is a technical description that defines this corpus and their associated contents.

[5] Rivest, R., "The MD5 Message-Digest Algorithm", RFC 1321, DOI 10.17487/RFC1321, April 1992, https://www.rfc-editor.org/info/rfc1321.

History

The (version) history of this corpus can be queried by:

```
preston history\
 --algo md5\
 --anchor hash://md5/53e144641ffded6800dea502a8bb47ed\
 --remotes https://linker.bio,https://zenodo.org
```

to produce:

```
<hash://md5/53e144641ffded6800dea502a8bb47ed> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/b855da1ec4dd0600c8fd736a0f413ec2> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/7290c3e847ed1d10e47e4fd12188b752> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/3e437559b980d2e88daf3c05ec2f4d77> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/352fa9f7e46576462cf7840b2992de0d> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/ac44c9ac6746a9fe2babaac45ec58860> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/d86a121f36f8e6f081a20803444bc1aa> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/1fb70a2c683eb7a5545900bc6668950f> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/63af62d86d0f3d4d26bfd534bf7282bb> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/b9014b0b403516b1e852748d72633c7d> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/f206133d0b42fc9126e36a30654748ec> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/98fbf06892fcf919bc2107446147cc4a> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/a28a615337c2c187a5d3286923642ad6> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/e19ad69093efab2b86c3ca04de8f95ba> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/feb6470b553aa9e104a6d7114a424a05> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/a445e6eb1138710be4b93db45ab7b4a9> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/016c9f11f52e6ce67c7399a4d588c2eb> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/a250d215282393ad43494dd9de11da4a> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<hash://md5/0502e4ea67c725e22f71aef16b164303> <http://www.w3.org/ns/prov#wasDerivedFrom> <ha
<urn:uuid:0659a54f-b713-4f86-a917-5be166a14110> <http://purl.org/pav/hasVersion> <hash://md5
```

Volume

The total volume of this corpus can be estimated using

```
preston ls\
 --algo md5\
 --anchor hash://md5/53e144641ffded6800dea502a8bb47ed\
 --remotes https://linker.bio,https://zenodo.org\
 | preston cat\
 | pv\
 > /dev/null
```

to produce:

```
190GiB
```

which suggest that the corpus entails a little under 200 GiB of data.

Also, the number of pdfs associated with content referenced in the corpus can be queried using

```
preston ls\
 --algo md5\
 --anchor hash://md5/53e144641ffded6800dea502a8bb47ed\
 --remotes https://linker.bio,https://zenodo.org\
 | grep hasVersion\
 | grep partpdf\
 | grep -v well-known\
 | wc -l
```

to produce 32320 items, or over 30k tracked pdfs. Note also that out of a total of 32739 items, pdf content related to 419 items failed to be retrieved. This number can be estimated using:

```
preston ls\
 --algo md5\
 --anchor hash://md5/53e144641ffded6800dea502a8bb47ed\
 --remotes https://linker.bio,https://zenodo.org\
 | grep hasVersion\
 | grep partpdf\
 | grep well-known\
 | wc -l
```