Unleashing Digital Knowledge into the Future

(a working document)

Jorrit H. Poelen

2023-10-11

[edit this page](https://github.com/bio-guoda/preston-service/edit/main/linker.bio/index.md) / [share suggestions](https://github.com/bio-guoda/preston/issues/new)

## Welcome to linker.bio

The internet is a powerful tool for exchanging digital information. But the Internet’s contents changes constantly: websites are launched and taken down, webpages change, and content gets archived or lost.

By design, a web address, or *U*niform *R*esource *L*ocator (URL), points to a specific internet *location* from which a resource, like a webpage, can be retrieved. However, a URL does *not* provide a way to verify that a retrieved webpage was the one we asked for. [[1]](#footnote-22)

Imagine using a URL-like reference to find a book at a library: instead of locating a book by what it is (e.g., title, author), you refer to a book by its location (e.g., third shelf on the second row next to the window). With this, a book becomes unfindable if moved to another shelf. And, if you do manage to find a book at the referenced location, how would you know you’ve found the book you are looking for?

Instead of pointing to where books are located, librarians point to them using a bibliographic reference. For practical reasons, only a few identifying clues are included in such a reference (e.g., author, year of publication, title, and publisher). So, librarians *refer to content by what it is*, and knowing where it may be located is secondary.

A bibliographic citation:

Darwin, C. 1859. On the Origin of Species. John Murray.

Thanks to recent advances in mathematics [[2]](#footnote-23), we can add digital fingerprints to bibliographic citations of *digital* content. A digital fingerprint uniquely describes any digital content (e.g., a webpage, an digital image, a pdf document) by a fixed length sequence of numbers and letters [[3]](#footnote-25). It is generated by performing a calculation [[4]](#footnote-26) on the content itself. Citations that include a digital fingerprint are also referred to as signed citations [[5]](#footnote-27).

These digital fingerprints open up a way to *automatically* verify, with astronomical certainty, that you got the digital content that you asked for.

## Asking for What You Want

On the internet, we’ve learned to say:

“I’d like to get the latest content from this web address.”

, and we *trust* that the retrieved content is what we asked for.

This may work well for a current news website or an internet search engine.

However, for retrieving specific content, like a newspaper article or research paper, we’d like to have a way of saying:

“I’d like to get the content with *this digital fingerprint*.”

, and *verify* that the retrieved content is *exactly* what we asked for.

In using digital fingerprints to refer to content, the location, and storage medium, of the content becomes secondary. In other words, *digital fingerprints help preserve references to digital content to a future beyond the internet*.

## The Content-Verse [[6]](#footnote-32)

This is where I’d like to introduce **the content-verse** as the collection of every single piece of digital content and each associated digital fingerprint. By definition, the content-verse contains all content (or knowledge) ever to be created.

## Using Digital Fingerprints

https://linker.bio builds a bridge [[7]](#footnote-34) from the exciting, dynamic internet to its reliable, boring, cousin — the **content-verse**. In this content-verse, **digital fingerprints are used as links** instead of those pesky, unreliable URLs. Contrary to URLs, these digital fingerprints **do *not* break, or expire** [[8]](#footnote-41).

Where the internet excels in spreading new information, the content-verse excels at referencing *known* information.

Through digital fingerprints, linker.bio provides a bridge to access billions of openly available biodiversity data records [[9]](#footnote-42), millions of Open Science publications through [Zenodo](https://zenodo.org), over eight hundred thousand datasets via [DataOne](https://dataone.org), billions of open source files [Software Heritage Library](https://softwareheritage.org), and more than ninety seven million freely usable media files from [WikiMedia Commons](https://commons.wikimedia.org/).

The beauty of digital fingerprints is that in fifty years from now, you may use that same fingerprint to find that information, regardless where it may be located, or how it is stored, or transmitted [[10]](#footnote-53).

## Requesting Content

https://linker.bio/ helps to request information, wherever it may be, using a notation like:

https://linker.bio/[fingerprint][.extension]

The extension is optional.

## Exploring Examples

For instance, to get a copy of a **scientific paper**, you can ask for:

[https://linker.bio/hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d.pdf](https://linker.bio/hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d).

or, to get a copy of a **scientific dataset**, like a historical CO2 Record from the Vostok Ice Core, you can ask for:

<https://linker.bio/hash://md5/e27c99a7f701dab97b7d09c467acf468>

or, perhaps even better, you can also ask for a **picture of a 🐇** (*Oryctolagus cuniculus*) by JM Ligero Loarte -

[https://linker.bio/hash://sha1/86fa30f32d9c557ea5d2a768e9c3595d3abb17a2.jpg](https://linker.bio/hash://sha1/86fa30f32d9c557ea5d2a768e9c3595d3abb17a2).

or, to review an initial draft of **the Hash URI Specification by Ben Trask** -

<https://linker.bio/hash://sha256/3fee21854fb6d81573b166c833db2771b21f0c77daa3095aab542764d89c94c1>.

## Building Your Own Bridge

So far, https://linker.bio appears to be a [“black box”](https://en.wikipedia.org/wiki/Black_box): you ask for some content by their fingerprint, and linker.bio attempts to retrieve that content.

But, what happens inside the box?

linker.bio is powered by [Preston](https://github.com/bio-guoda/preston). Preston builds a bridge from the content-verse (e.g., a digital fingerprint) to the content stored in physical locations. Preston is the little machine that responds when you ask for the picture of the bunny using the URL https://linker.bio/hash://sha1/86fa30f32d9c557ea5d2a768e9c3595d3abb17a2. And, if you know how to run a program on your computer, you can run your own machine (or server) that looks up that bunny picture. In other words, with a little effort, you can build your own bridge without have to ask for permission or paying some kind of license fee[[11]](#footnote-63).

For the tech savvy, you can run Preston in server mode on linux/mac [[12]](#footnote-64) by executing the following in the terminal:

preston server --remote https://wikimedia.org

or,

preston s --remote https://wikimedia.org

for short.

On starting the server, you’ll see some cryptic messages that end with

[main] INFO org.eclipse.jetty.server.AbstractConnector - Started ServerConnector@76a4d6c{HTTP/1.1, (http/1.1)}{localhost:8080}  
[main] INFO org.eclipse.jetty.server.Server - Started @561ms

This means that the Preston server is waiting for requests.

Now, you can visit http://localhost:8080/hash://sha1/86fa30f32d9c557ea5d2a768e9c3595d3abb17a2 to retrieve the bunny picture. On receiving your question, Preston will try and ask https://wikimedia.org whether is has any content in their https://commons.wikimedia.org/ library with that digital fingerprint. If so, Preston will ask Wikimedia to send that content, and then, pass it to you. The next time you ask for the bunny picture, you’ll receive the picture pretty fast, because Preston remembers the content associated to the digital fingerprint and doesn’t have to ask https://wikimedia.org anymore.

In addition to Wikimedia Commons, Preston knows how to talk to https://zenodo.org, https://softwareheritage.org, https://dataone.org and … other Preston servers!

So, in the previous example, Preston talked directly to wikimedia.org . In the example below, your Preston server would talk to https://linker.bio instead, and https://linker.bio would relay the request:

preston s --remote https://linker.bio

You can even provide a list of “remotes”. If a list is provided, Preston asks the provided location in order of appearance. With the example below, Preston would first ask linker.bio, then if linker.bio doesn’t have it, it’ll ask wikimedia.org.

preston s --remote https://linker.bio,https://wikimedia.org

So, with this you can create elaborate combinations of ways to ask for content. One example of such elaborate setup is a [content delivery network](https://en.wikipedia.org/wiki/Content_delivery_network) to facilitate reliable access to well-known content.

If you’d like to learn more about how to run a Preston server, but don’t know where to start, please send an email to [Jorrit](mailto:jhpoelen+preston@jhpoelen.nl) or open a [GitHub issue](https://github.com/bio-guoda/preston/issues/new).

## A Use Case: Studying Pine Pests Caused by Weevils (Curculionoidea)

Imagine studying a pine pest caused by weevils, plant eating beetles of super order Curculionoidea. In preparation for answering a research question, you may want to understand what is known about them, and some of their hosts pine trees. By combining large versioned corpora compiled using digital fingerprints, you can answer complex questions across disciplines. The examples below show how some related questions span digital collections made available through natural history collections, taxonomic literature, genetic records of plants, and biodiversity literature.

By using nimble, yet powerful, data processing tools like Preston and Nomer to make your laptop (or some powerful computer you have access to) to answer complex and specific *interdiscplinary* research questions.

| corpus | last updated | size (approx) | version |
| --- | --- | --- | --- |
| GIB (**G**BIF, **i**DigBio and **B**ioCASe) | 2023-09-01 | ~2.7TB | [hash://sha256/a755a6…](https://linker.bio/hash://sha256/a755a6ac881e977bc32f11536672bfb347cf1b7657446a8a699abb639de59419) |
| ChecklistBank | 2023-09-01 | ~30GB | [hash://sha256/8cb357…](https://linker.bio/hash://sha256/8cb3570748849866dc59f72aa6711e9a10111049063948fd0742f0ea2aeb6290) |
| Biodiversity Heritage Library | 2023-09-01 | ~300GB | [hash://sha256/9cc132…](https://linker.bio/hash://sha256/9cc1327e220df2c215b716a8ab07fcd9e3bffc1f6a808dbb6957f40e3df8c94c) |
| GenBank PLN Division | 2023-06-28 | ~250GB | [hash://sha256/efa589…](https://linker.bio/hash://sha256/efa589e9927c8c10ac81867220982a3eb732c322ddebca6db2a9147578218ff5) |
| Nomer Corpus of Taxonomic Resources | 2023-09-08 | 10GB | [hash://sha256/12051b…](https://linker.bio/hash://sha256/12051b8aa59930d6561a3ed46b7cf3f67a31a98445a457d78894f6b8a8e81641) |
| [OpenAlex](https://openalex.org)[[13]](#footnote-74) | 2023-11-01 | ~300GB | [hash://sha256/f19011…](https://linker.bio/hash://sha256/f19011fe72234cb22f4326fcc95752647a11628cdaa0d24f7d55033b06cb5653)[[14]](#footnote-76) |

To prepare for addressing more complex research question, the following basic question may need answering:

Q1. How many specimen of Weevils (plant eating beetles of super order Curculionoidea) have been recorded globally?

The following steps can help towards answering Q1.

step 1. list GIB corpus content at version a755...  
step 2. print all related biodiversity records  
step 3. for each record, select origin and scientific name  
step 4. align names with Catalogue of Life as included in Nomer Corpus version 1205...  
step 5. count only aligned records that mention "Curculionoidea"

And these steps may be implemented in linux bash using Preston, jq, mlr, Nomer, and grep:

preston cat\  
 --no-cache\  
 --remote https://linker.bio\  
 hash://sha256/a755a6ac881e977bc32f11536672bfb347cf1b7657446a8a699abb639de59419\  
 | preston dwc-stream\  
 --no-cache\  
 --remote https://linker.bio\  
 | jq -c '{ "src": .["http://www.w3.org/ns/prov#wasDerivedFrom"], "name": .["http://rs.tdwg.org/dwc/terms/scientificName"] }'\  
 | mlr --ijsonl --otsv cat\  
 | nomer append col\  
 | grep -v NONE\  
 | grep Curculionoidea\  
 | pv -l

where lines 1-4 implement step 1., lines 5-7 implement step 2., lines 8-9 implement step 3., line 10 implements step 4., and lines 11-13 implement step 5.

The implementation above shows a brute force way to answer Q1. Resources permitting, optimized workflows can be generated to allow for more responsive and specialized services. So, you can answer your question based on versioned digital corpora, and optimize when needed / possible. One such optimization can be to clone the needed versioned datasets on a fast solid state drive to reduce network delays. Another could be to configure a search engine to help answer a selective kind of user queries quickly. And, with the versioned corpus, these uses of the datasets of known origin can be built independently of the corpus itself, allowing for teams work independently to improve the use of a well-defined knowledge corpus.

Other possible questions include:

Q2. How many distinct species of Weevils, or plant eating beetles of the superfamily Curculionoidea have been described in taxonomic literature and checklists?

Q3. How time do species names of Weevils (plant eating beetles of the superfamily Curculionoidea) occur in the a copy of transcribed texts made available through the Biodiversity Heritage Library? Q4. How many genetic sequences are available for Pinus taeda (loblolly pine) are available through GenBank?

For more information and background, see:

Elliott, M.J., Poelen, J.H. & Fortes, J.A.B. Signing data citations enables data verification and citation persistence. Sci Data 10, 419 (2023). <https://doi.org/10.1038/s41597-023-02230-y> [hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d](https://linker.bio/hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d.pdf)

## Acknowledgments

Michael Elliott, José Fortes and [Cypress Hansen](https://www.cypresswritesscience.com/) provided comments to help improve a description of todays internet and the benefits of the content-verse.

1. A similar argument can made for a *D*igital *O*bject *I*dentifier (DOI, https://en.wikipedia.org/wiki/Digital\_object\_identifier). [↑](#footnote-ref-22)
2. Sobti, R. & Geetha, G. Cryptographic Hash Functions: A Review. 2012. International Journal of Computer Science Issues (IJCSI) 9, 461–479 <https://www.ijcsi.org/papers/IJCSI-9-2-2-461-479.pdf> accessed at 2023-10-11 with hash://md5/eb8e2fb3e16bd5839443cd40a9a8c3c1 [↑](#footnote-ref-23)
3. The length of a fingerprint depends on the kind of calculation used, and typically vary between 32 (for MD5), 40 (for sha1) or 64 (for SHA256) characters. [↑](#footnote-ref-25)
4. Such calculations are referred to as cryptographic hash functions like MD5, or SHA-256 [↑](#footnote-ref-26)
5. Elliott, M.J., Poelen, J.H. & Fortes, J.A.B. Signing data citations enables data verification and citation persistence. Sci Data 10, 419 (2023). <https://doi.org/10.1038/s41597-023-02230-y> [hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d](https://linker.bio/hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d.pdf) [↑](#footnote-ref-27)
6. Probably need a better term for this, because of existing uses elsewhere. [↑](#footnote-ref-32)
7. linker.bio is not the only bridge to the content-verse. In fact, linker.bio re-uses existing bridges provided by [Zenodo](https://github.com/bio-guoda/preston/issues/149), [DataOne](https://github.com/bio-guoda/preston/issues/181), [WikiMedia Commons](https://github.com/bio-guoda/preston/issues/239), and [Software Heritage Library](https://github.com/bio-guoda/preston/issues/70) to the massive amount of content they keep. Note also that [Carl Boettiger](https://www.carlboettiger.info/) maintains a bridge from R to the content-verse via the R package [contentid](https://github.com/cboettig/contentid). [↑](#footnote-ref-34)
8. Digital finger prints are cryptographic objects that are mathematically linked to the content they reference. They can be generated with most digital devices. And, by embedding fingerprints into other digital content, you can pretty much reference anything and everything digital using a digital fingerprint that fits on a T-shirt. [↑](#footnote-ref-41)
9. Biodiversity records include snapshot version of digital collections registered with [iDigBio](https://idigbio.org), [GBIF](https://gbif.org), [BioCase](https://biocase.org), [Biodiversity Heritage Library](https://biodiversitylibrary.org), [OBIS](https://obis.org) and [CheckListbank](https://checklistbank.org) [↑](#footnote-ref-42)
10. In other words, digital fingerprints are agnostic of location, technology, and … time. [↑](#footnote-ref-53)
11. Preston is open source software. [↑](#footnote-ref-63)
12. Windows is supported too, but you’ll have to run Preston a little differently. See documentation for examples. [↑](#footnote-ref-64)
13. [OpenAlex](https://openalex.org) is an open access index of scientific works, similar to Web of Science and Google Scholar but then without paywalls or walled gardens. [↑](#footnote-ref-74)
14. See <https://gist.github.com/jhpoelen/8b263027ff13c7b788fa24866ce73bfb> for example usage of a versioned copy of the OpenAlex index. [↑](#footnote-ref-76)