A DataVerse Beyond The Internet

Jorrit Poelen https://jhpoelen.nl

2024-02-06

Guiding Questions

- How do *you* cite data?
- How do you look up cited data *now*?
- How do you look up cited data *40 years from now*?

There's A Kitty In The DataVerse...

Figure 1: Harvard Kitty

There's A Kitty In The DataVerse...

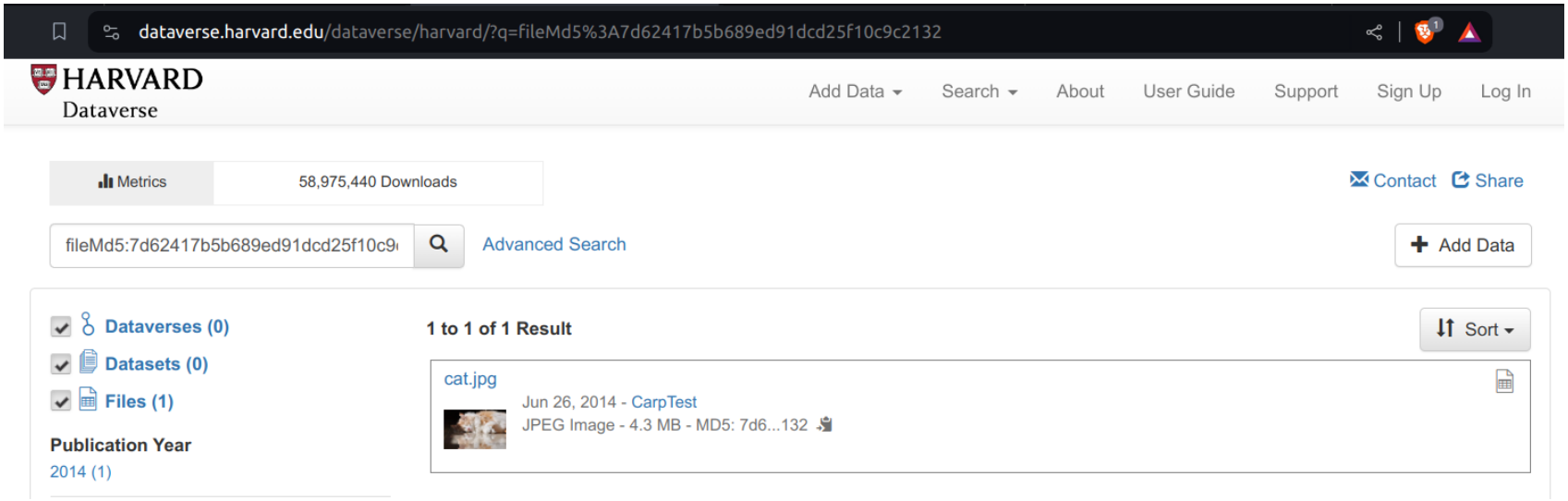


Figure 2: Harvard Kitty Internet Page

So How'd You Cite This Kitty Picture?

Joshua Carp, 2014, "cat.jpg", CarpTest, https://doi.org/10.7910/DVN/24358/N4FCVS, Harvard Dataverse, V1 [1]

[1] As suggested in https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/24358/N4FCVS accessed on 2024-02-06 https://dataverse.org/best-practices/data-citation

How To Retrieve This Kitty Picture Now?

Like this?

```
curl -L "https://doi.org/10.7910/DVN/24358/N4FCVS"\
 > cat.jpg
```

How To Retrieve This Kitty Picture Now?

Wait a minute . . .

```
cat cat.jpg
 | head -n2
```

```
<?xml version='1.0' encoding='UTF-8' ?>
<!DOCTYPE html>
```

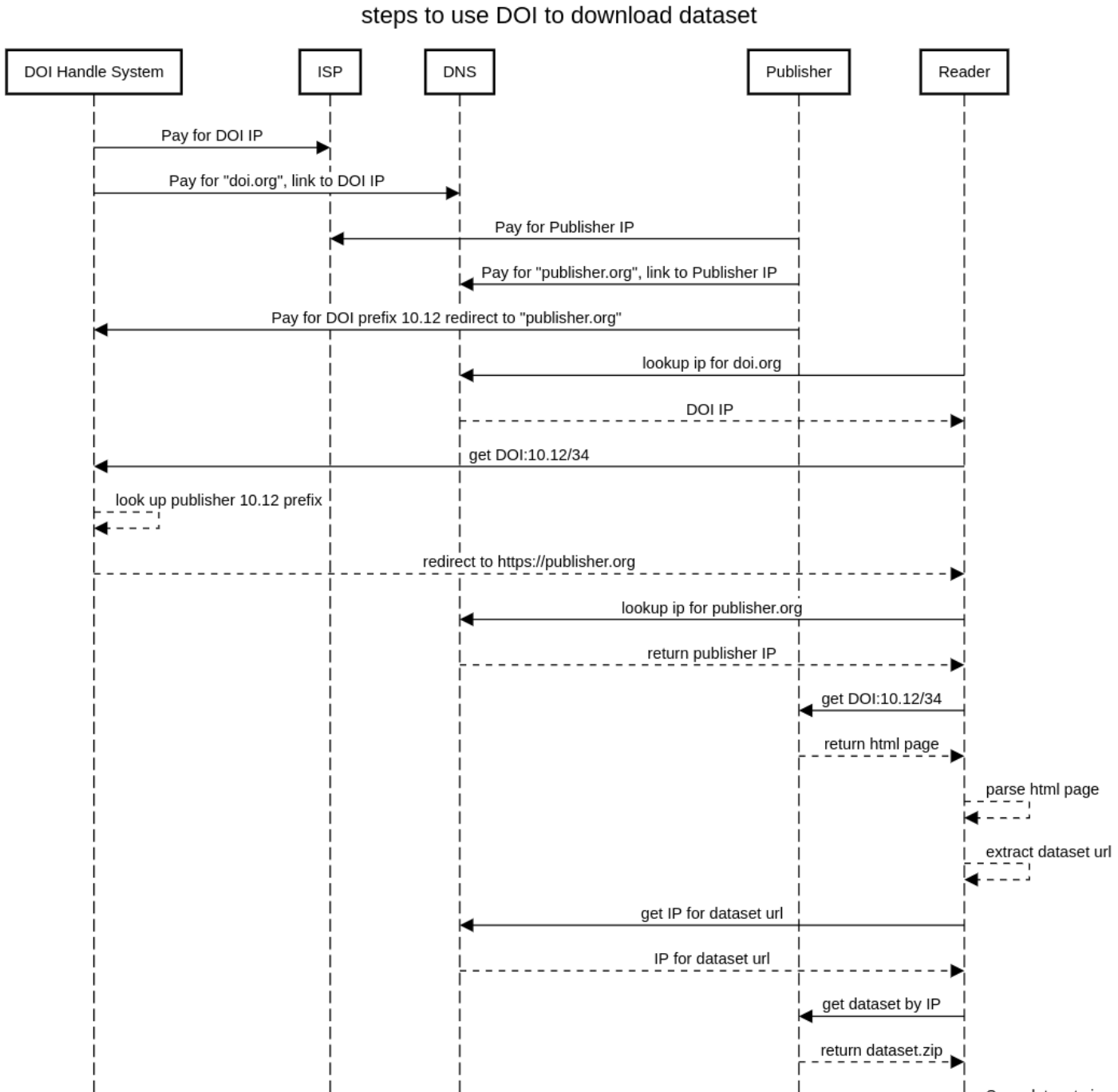
That ain't no cat, it is an HTML page.

How To Retrieve This Cat Picture Now?

Take 2.

- turn on the internet (still there?)
- open a web browser
- load https://doi.org/10.7910/DVN/24358/N4FCVS
  - and rely on a delicate and complex socio-technical network
- inspect page
- use mouse to click on a link that looks like cat.jpg
- download image and **trust** its authenticity

DOI Economics and Redirection



How To Retrieve This Cat Picture 50 Years From Now?

Joshua Carp, 2014, "cat.jpg", CarpTest, https://doi.org/10.7910/DVN/24358/N4FCVS, Harvard Dataverse, V1

Likely will not work due to intricate network of dependencies.

How To Retrieve This Cat Picture 50 Years From Now?

Proposal: Sign the citation [2]

. . . by adding the digital fingerprint of the image.

Joshua Carp, 2014, "cat.jpg", CarpTest, https://doi.org/10.7910/DVN/24358/N4FCVS, Harvard Dataverse, V1
hash://md5/7d62417b5b689ed91dcd25f10c9c2132

---

[2] 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

How To Retrieve This Cat Picture 50 Years From Now?

. . . by searching the world (in, and beyond, the internet) for the content with the unique fingerprint hash://md5/7d62417b5b689ed91dcd25f10c9c2132 .

- https://linker.bio/hash://md5/7d62417b5b689ed91dcd25f10c9c2132
- https://dataverse.harvard.edu/dataverse/harvard/?q=fileMd5%3A7d6
- `preston cat --remote https://dataverse.harvard.edu hash://md5/7d62417b5b689ed91dcd25f10c9c2132`
- `preston cat --remote https://dataverse.org hash://md5/7d62417b5b689ed91dcd25f10c9c2132`
- `preston cat --remote https://linker.bio,https://databverse.org hash://md5/7d62417b5b689ed91dcd25f10c9c2132`

Internet Is Designed For *Exchanging* Information

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.[3]

[3] Jorrit Poelen. 2024. Unleashing Digital Knowledge Into The Future. Accessed on 2024-02-06 at https://linker.bio line:hash://sha256/8ac18eb75ff20d40d1d60bb6ad5a745eb528093d1ffbe373e3847c131460

Internet Is *Location-based*

By design, a web address, or Uniform Resource Locator (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. [4]

[4] Jorrit Poelen. 2024. Unleashing Digital Knowledge Into The Future. Accessed on 2024-02-06 at https://linker.bio line:hash://sha256/8ac18eb75ff20d40d1d60bb6ad5a745eb528093d1ffbe373e3847c131460

Finding Content By Their Location Is . . . Tricky

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? [5]

[5] Jorrit Poelen. 2024. Unleashing Digital Knowledge Into The Future. Accessed on 2024-02-06 at https://linker.bio line:hash://sha256/8ac18eb75ff20d40d1d60bb6ad5a745eb528093d1ffbe373e3847c131460

Finding Content By Their (Summarized) Content Is . . . What Librarians Do

> *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.*** [6]

A bibliographic citation:

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

[6] Jorrit Poelen. 2024. Unleashing Digital Knowledge Into The Future. Accessed on 2024-02-06 at https://linker.bio line:hash://sha256/8ac18eb75ff20d40d1d60bb6ad5a745eb528093d1ffbe373e3847c131460

Turns out that there are cats in the DataVerse too . . .

```
preston cat\
 --remote "https://dataverse.org"\
 hash://md5/7d62417b5b689ed91dcd25f10c9c2132\
 > cat.jpg
```