

# Tools for Digital Data Recovery

Naomi Alterman // May 19 2025  
eScience Institute



[http://staff.uw.edu/naomila/slides/  
2025-data-recovery-panel](http://staff.uw.edu/naomila/slides/2025-data-recovery-panel)

*“Epimetheus opening Pandora's box”*  
Giulio Bonasone [[source](#)]



## Two separate problems:

1. Identifying data to retrieve

2. Capturing identified data



# Taxonomy of data tools

# How is your data accessed?

*(from the perspective of a software engineer)*

- Web browser  
(articles, interactive exhibits, streaming video, ...)
- Downloaded files  
(PDFs, spreadsheets, hi-res images, ...)
- Queried content  
(library catalogue metadata, databases, earth-scale imaging, ...)



Website archiving  
tools



Bulk downloading  
tools



Query languages  
and APIs

# How is your data accessed?

*(from the perspective of a software engineer)*

- Web browser  
(articles, interactive exhibits, streaming video, ...)



Website archiving  
tools

- Downloaded files  
(PDFs, spreadsheets, hi-res images, ...)



Bulk downloading  
tools

- Queried content  
(library catalogue metadata, databases, earth-scale  
imaging, ...)



Query languages  
and APIs

# Web archives

- Tools that capture web pages and their constituent media into a “**frozen snapshot**” of a site at a point in time
  - In the spirit of <https://web.archive.org/>
  - Both **automatic** and **manual** variants of these tools
  - Output of the process: “**WARC**” files
- Resources to learn more:
  - <https://github.com/iipc/awesome-web-archiving?tab=readme-ov-file>
  - [https://archive.org/details/introduction-to-the-warc\\_202111](https://archive.org/details/introduction-to-the-warc_202111)
  - <https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml>

http://www.spacejam.com/jam.htm

Go

DEC

JAN

FEB

17

[156 captures](#)


12 Apr 1997 - 17 Jun 2024

1998

1999

2000

About this capture

 [Click here](#)  
for the straight  
facts on marijuana.

**World  
Conquest** 



PRESS BOX SHUTTLE



JAM CENTRAL



PLANET B-BALL



LUNAR TUNES



THE LINEUP



JUNIOR JAM

STELLAR SOUVENIRS



JUMP STATION

WARNER STUDIO STORE



**SPACE JAM**



SITE MAP

 [Behind  
the Jam](#)

ADVERTISEMENT: [Click Here For Great Specials](#)

[Click here!](#)  
Win A  
BMW Z3

**Go**  
com

DVD is now at  
**TOTAL**  [click  
here](#)

[Find Food, Fun & more  
in your  
own Backyard!](#)

Buy a gift for

**FREE TICKETS**

# How is your data accessed?

*(from the perspective of a software engineer)*

- Web browser  
(articles, interactive exhibits, streaming video, ...)



Website archiving  
tools

- Downloaded files  
(PDFs, spreadsheets, hi-res images, ...)



Bulk download tools,  
web scraping tools

- Queried content  
(library catalogue metadata, databases, earth-scale  
imaging, ...)



Query languages  
and APIs



# Bulk downloading

- Files on the internet are generally referenced by a **URL** (“uniform resource locator”)
- Our goal for downloading a corpus of papers is to **collect a list** of **URLs** for each paper, and then use a **download manager** to retrieve them
- Often: we use a scripting language like **Python** or **R** to collect said URLs from a catalogue

# Download managers

- MacOS / Linux “power user” tool: **wget**  
<https://www.digitalocean.com/community/tutorials/how-to-use-wget-to-download-files-and-interact-with-rest-apis>
- Windows: graphical wget tool  
<https://winwget.sourceforge.net/>

# How is your data accessed?

*(from the perspective of a software engineer)*

- Web browser  
(articles, interactive exhibits, streaming video, ...)



Website archiving  
tools

- Downloaded files  
(PDFs, spreadsheets, hi-res images, ...)



Bulk download tools,  
web scraping tools

- Queried content  
(library catalogue metadata, databases, earth-scale  
imaging, ...)



Query languages  
and APIs

# Link collection through web scraping

- **Web scraping:** scripts use web scraping tools to extract information from human-oriented web pages
  - Easy mode: [scrapy](#), [portia](#), [beautiful soup](#)
  - Hard mode: [Selenium](#) and other “dynamic” web scrapers that can interact with page elements
- Resources:
  - <https://github.com/lorien/awesome-web-scraping/blob/master/python.md#web-scraping--frameworks>
  - For R: <https://datasciencebox.org/02-exploring-data#web-scraping-and-programming>

# Link collection through APIs

- “**Application Programming Interfaces**”: interfaces for code to directly communicate with a database or other information system, bypassing human-centric interfaces
  - Harder to learn, but ultimately better to use if given the opportunity
- Web APIs have URLs just like websites do, but they’re used for **information transactions** rather than accessing webpages
  - These transactions have a **request** (that we send) and a **response** (that we get back)
  - Information is often exchanged as **JSON** (pronounced “jay-son”) or **XML**
- An API **schema** is a rigorous specification of how API requests and responses should be structured

# OAI-PMH

- “Open Archives Initiative - Protocol for Metadata Harvesting”
  - <https://www.openarchives.org/pmh/>
- A common API schema to request library catalogue metadata (among other things)
  - Used by *many* government-hosted research repositories and stacks

# API tools

- General non-code tools:
  - Insomnia: <https://insomnia.rest/>
  - Curleroo: <https://app.curleroo.com/>
- General code tools:
  - Python “**requests**” library: <https://pypi.org/project/requests/>
- OAI-PMH specific:
  - Omeka plugin: <https://omeka.org/classic/docs/Plugins/OaipmhHarvester/>
  - Web tool: <https://validator.oaipmh.com/>
  - Python library: <https://github.com/infrae/pyoai>



**National Oceanic and  
Atmospheric Administration**  
United States Department of Commerce



[Advanced Search](#)

Home

Collections

Recent Additions

Submit

Submission  
Information

Help

About NOAA Inst. Repos.



## Search our Collections



[Advanced Search](#)

Search



### Deepwater Horizon Oil Spill and Restoration (DWH)

A collection of assessment/restoration documents pertaining to the 2010 Deepwater Horizon oil spill.

[Learn More](#)







## Help Center

What can we help you with?

### SEARCH TIPS:

[How to Perform an Simple Search](#)[How to Perform an Advanced Search](#)

### CITATION EXPORT

[How to Export Citation from Single Document](#)

### HARVEST METADATA

[OAI-PMH](#)[JSON/CSV/XML](#)

### FAQs

[Home](#)[Collections](#)[Recent Additions](#)[Coming Soon](#)[Search Results](#)[Facets](#)

**Not seeing what you need? Still need help?**

[email us at publishing@cdc.gov](mailto:publishing@cdc.gov)

Some groups will want to harvest metadata descriptions of records in an archive so that services can be built using metadata from many archives. This capability is implemented using two different options: the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and JSON.

**What is OAI-PMH?** The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol developed for harvesting metadata descriptions of records in an archive so that services can be built using metadata from many archives. An implementation of OAI-PMH must support representing metadata in Dublin Core, but may also support additional representations.

OAI-PMH specification is defined as a RESTful service, where each request requires certain attributes. A "verb" is described as to what is being asked. Each verb can have different attributes.

A full description of the specifications can be found [here](#).

OAI Verbs: Based on the 2.2 spec, the available verbs are: (the verbs are case sensitive): Identify, ListMetadataFormats, GetRecord, ListIdentifiers, ListRecords

Examples Requests (Dev):

- [https://\[stacksurl\]/fedora/oai?verb=Identify](https://[stacksurl]/fedora/oai?verb=Identify)
- [https://\[stacksurl\]/fedora/oai?verb=ListRecords&metadataPrefix=oai\\_dc&from=2017-05-01T00:00:00Z&until=2017-05-01T15:16:46Z](https://[stacksurl]/fedora/oai?verb=ListRecords&metadataPrefix=oai_dc&from=2017-05-01T00:00:00Z&until=2017-05-01T15:16:46Z)
- [https://\[stacksurl\]/fedora/oai?verb=GetRecord&metadataPrefix=oai\\_dc&identifier=oai:cdc.stacks:cdc:32147](https://[stacksurl]/fedora/oai?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:cdc.stacks:cdc:32147)
- [https://\[stacksurl\]/fedora/oai?verb=ListMetadataFormats](https://[stacksurl]/fedora/oai?verb=ListMetadataFormats)
- [https://\[stacksurl\]/fedora/oai?verb=ListSets](https://[stacksurl]/fedora/oai?verb=ListSets)
- [https://\[stacksurl\]/fedora/oai?verb=ListIdentifiers&metadataPrefix=oai\\_dc](https://[stacksurl]/fedora/oai?verb=ListIdentifiers&metadataPrefix=oai_dc)

Note: the resumptionToken parameter is mutually exclusive with other parameters. It is used to paginate through results and the token is given by each request under the tag at the end of the response.

Ex.: e80b606013a0a8274d91bb6b95d94bba

The first request will pass the appropriate values for other parameters, including metadataPrefix, from and until. subsequent pages have to pass only resumptionToken.

To test the resumptionToken through the browser, enter this URL:

# Insomnia NOAA-IR example

The screenshot displays the Insomnia REST client interface. At the top, a tab labeled 'GET' is active for the URL 'https://repository.library.noaa.gov/fedora/oai'. The status bar indicates a successful response with a 200 OK status, a response time of 1.1 seconds, and a body size of 242.5 KB, received 21 minutes ago.

The left sidebar shows the 'Params' tab with two query parameters:

Key	Value
verb	ListRecords
metadataPrefix	oai_dc

The 'URL PREVIEW' section shows the constructed URL: 'https://repository.library.noaa.gov/fedora/oai?verb=ListRecords&metadataPrefix=oai\_dc'.

The right pane shows the 'Preview' tab with the XML response body. The XML is an OAI-DC record for a document titled 'Pilot-Scale Production Economics Of C. Ariakensis Oysters : Summary Of Virginia Seafood Council Industry Grow- Out Trials (2003-2005)'. The identifier is 'https://repository.library.noaa.gov/view/noaa/38420'. The subjects are 'Oyster culture', 'Economic aspects', and 'Suminoe oyster'. The description details the Virginia Fishery Resource Grant (VFRGP) program.

```
<?xml version="1.0" encoding="UTF-8"?>
<header>
  <identifier>oai: NOAA.stacks: NOAA:38420</identifier>
  <timestamp>2025-05-14T17:38:23Z</timestamp>
</header>
<metadata>
  <oai_dc:dc
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
    <dc:title>Pilot-Scale Production Economics Of C. Ariakensis Oysters :
Summary Of Virginia Seafood Council Industry Grow- Out Trials (2003-2005)
</dc:title>
    <dc:title>Pilot-Scale Production Economics Of "C.
Ariakensis" Oysters: Summary Of Virginia Seafood Council Industry Grow- Out
Trials (2003-2005)</dc:title>
    <dc:identifier>https://repository.library.noaa.gov/view/noaa/38420</dc:identifie
r>
    <dc:subject>Oyster culture</dc:subject>
    <dc:subject>Economic aspects</dc:subject>
    <dc:subject>Suminoe oyster</dc:subject>
    <dc:description>The "Virginia Fishery Resource Grant
Program" (VFRGP) was initiated by the Virginia Legislature to "protect and
enhance the Commonwealth's coastal fishery resource through the
awarding of grants in four areas": 1)New fisheries equipment or gear;
2)Environmental pilot studies on issues including water quality and fisheries
habitat; 3)Aquaculture or mariculture of marine-dependent species; and
4)Seafood technology. The VFRGP is based on the simple approach that experienced
fishermen can develop effective ideas for improving productivity or reducing
costs. Typically, attempting such an idea or change entails a cash outlay that
```

Thinking like an engineer

# Think like an engineer

- We might need to stitch several data retrieval methods together
  - Example: API only provides catalog identifiers and abstracts for papers, but not the papers themselves
  - This doesn't mean the data provider doesn't want us to bulk export! It just means we need to **engineer** a solution that uses several tools

# Data retrieval sketch

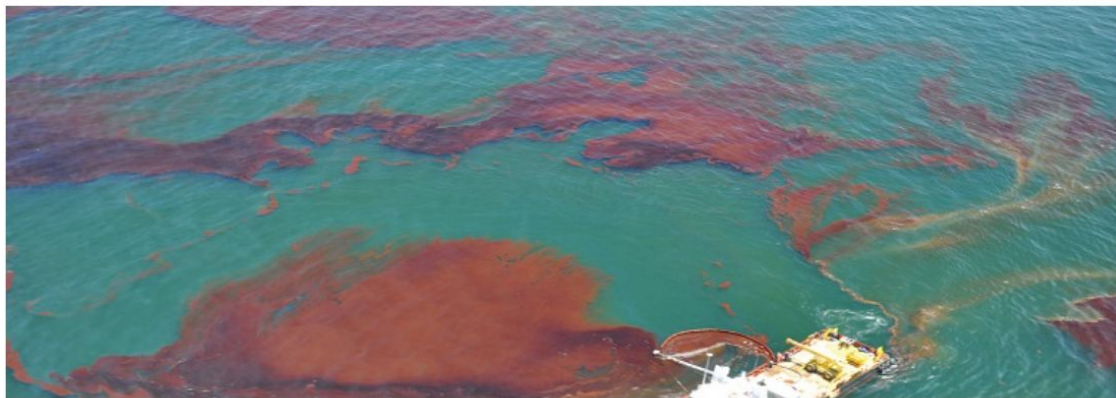
- **Python script** that uses OAI-PMH **API** to **identify the DOIs** of papers we want to save in the NOAA-IR stacks
  - **Assistive technologies:** Python “**requests**” library
  - **Input:** N/A, **Output:** list of DOIs
- **Python script** that, for each DOI, loads the human-facing catalog page and **web scrapes** the paper’s download link from it
  - **Assistive technologies:** Python “**beautifulsoup**” library
  - **Input:** List of document DOIs, **Output:** list of PDF download links
- Use WinWGet **download manager** to **bulk-download** PDFs in parallel
  - **Input:** List of PDF download links, **Output:** Final archival materials



**National Oceanic and  
Atmospheric Administration**  
United States Department of Commerce

[Advanced Search](#)[Home](#)[Collections](#)[Recent Additions](#)[Submit](#)[Submission  
Information](#)[Help](#)[About NOAA Inst. Repos. ▾](#)

## Search our Collections

[Advanced Search](#)[Search](#)

### Deepwater Horizon Oil Spill and Restoration (DWH)

A collection of assessment/restoration documents pertaining to the 2010 Deepwater Horizon oil spill.

[Learn More](#)



# Insomnia NOAA-IR example

Scratch Pad × GET New Request × +

GET https://repository.library.noaa.gov/fedora/oai Send 200 OK 1.1 s 242.5 KB 21 Minutes Ago

Params 2 Body Auth Headers 3 Scripts Docs

URL PREVIEW

https://repository.library.noaa.gov/fedora/oai?verb=ListRecords&metadataPrefix=oai\_dc

QUERY PARAMETERS Import from URL Bulk Edit

+ Add Delete all Description

verb	ListRecords	✓	✕
metadataPrefix	oai_dc	✓	✕

Preview Headers 14 Cookies Tests 0/0 → Mock Console

Preview

```
9 <header>
10   <identifier>oai:noaa.stacks:noaa:38420</identifier>
11   <timestamp>2025-05-14T17:38:23Z</timestamp>
12 </header>
13 <metadata>
14   <oai_dc:dc
15     xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
16     xmlns:dc="http://purl.org/dc/elements/1.1/"
17     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
18     xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
19       http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
20     <dc:title>Pilot-Scale Production Economics Of C. Ariakensis Oysters :
21       Summary Of Virginia Seafood Council Industry Grow- Out Trials (2003-2005)
22     </dc:title>
23     <dc:title>Pilot-Scale Production Economics Of "C.
24       Ariakensis" Oysters, Summary Of Virginia Seafood Council Industry Grow- Out
25       Trials (2003-2005) </dc:title>
26   <dc:identifier>https://repository.library.noaa.gov/view/noaa/38420</dc:identifier>
27   <dc:subject>Oyster culture</dc:subject>
28   <dc:subject>Economic aspects of</dc:subject>
29   <dc:subject>Suminoe oyster</dc:subject>
30   <dc:description>The "Virginia Fishery Resource Grant
31     Program" (VFRGP) was initiated by the Virginia Legislature to "protect and
32     enhance the Commonwealth's coastal fishery resource through the
33     awarding of grants in four areas": 1)New fisheries equipment or gear;
34     2)Environmental pilot studies on issues including water quality and fisheries
35     habitat; 3)Aquaculture or mariculture of marine-dependent species; and
36     4)Seafood technology. The VFRGP is based on the simple approach that experienced
37     fishermen can develop effective ideas for improving productivity or reducing
38     costs. Typically, attempting such an idea or change entails a cash outlay that
```





The NOAA IR serves as an archival repository of NOAA-published products including scientific findings, journal articles, guidelines, recommendations, or other information authored or co-authored by NOAA or funded partners. As a repository, the NOAA IR retains documents in their original published format to ensure public access to scientific information.

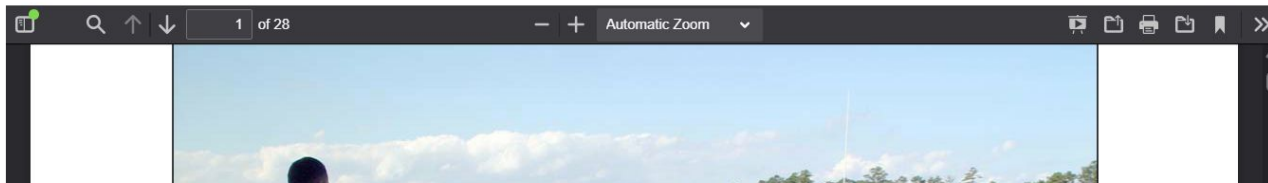


## **Pilot-Scale Production Economics Of *C. Ariakensis* Oysters : Summary Of Virginia Seafood Council Industry Grow- Out Trials (2003-2005)**

2005

By [Murray, Thomas J.](#)

Series: [VIMS Marine Resource Report](#) ; 2005-04



[Download Document](#)

[CITE](#)



Place as Subject:

Virginia

License:

<https://creativecommons.org/publicdomain/zero/1.0/>

Rights Information:

Public Domain

Compliance:

Library

Main Document Checksum:

[+]

Download URL:

[https://repository.library.noaa.gov/view/noaa/38420/noaa\\_38420\\_DS1.pdf](https://repository.library.noaa.gov/view/noaa/38420/noaa_38420_DS1.pdf)

File Type:



[PDF-506.11 KB]

## You May Also Like

Non-technical considerations and  
further resources

# Relationship to data provider

- **Collaborative**

- Provides bulk data request tools, FTP access, provides open and documented APIs, ...







- **Neglectful**

- Hasn't seemed to think or worry about data export at all, possibly provides APIs but they don't quite cover what we need, ...

- **Adversarial**

- Has a vested interest in preventing data export, monitors and prevents data export attempts, ...

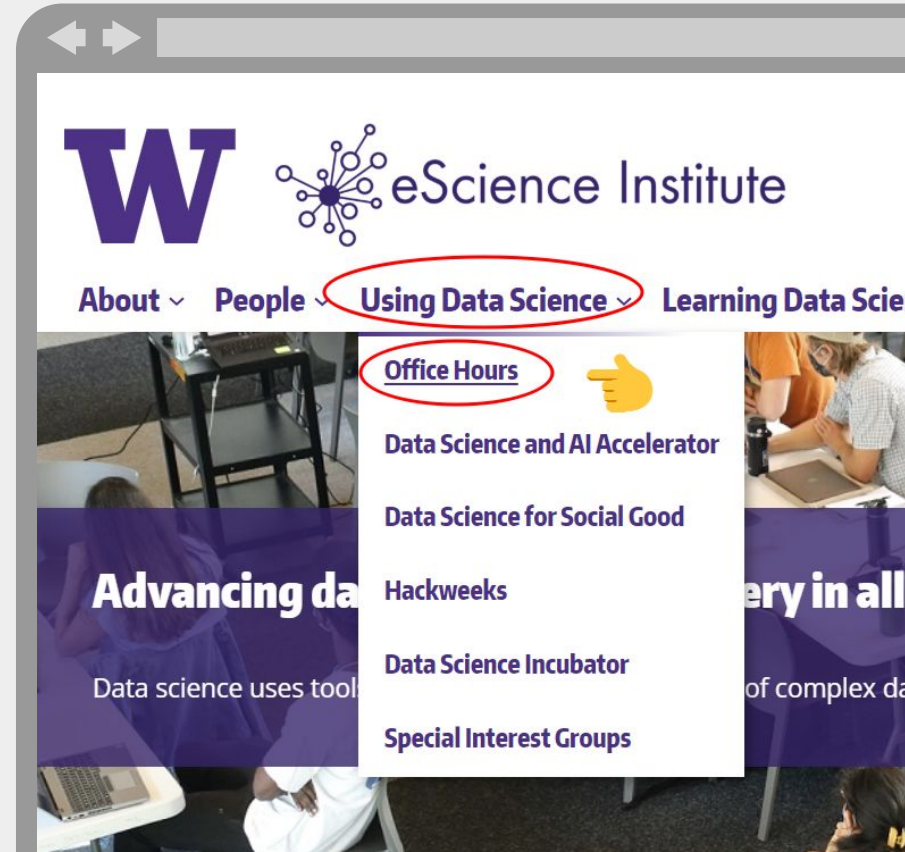
# A word about chat-based AI tools

- Using ChatGPT to help you construct API requests or write Python scripts?
  -   
  - These outputs are inspectable, reproducible and verifiable.
- Using ChatGPT to directly aggregate download links from a list of webpages or retrieve metadata?
  -   
  - These outputs are opaque and not actually produced via rigorous procedure, even if they claim to be

# eScience Office Hours

<https://escience.uw.edu/oh>

By-appointment consultation, help and planning for data archival projects (among other things)





[staff.uw.edu/naomila/slides/  
2025-data-recovery-panel](https://staff.uw.edu/naomila/slides/2025-data-recovery-panel)