

Big Data: Data Wrangling Boot Camp  
Web Crawling with R and OAI-PMH

Chuck Cartledge, PhD

25 February 2018

# Table of contents (1 of 1)

- 1 Intro.
- 2 OAI-PMH
  - What is OAI-PMH
  - NASA Reports
- 3 Hands-on
- 4 Q & A
- 5 Conclusion
- 6 References
- 7 Files

# What are we going to cover?

- Look to the future
- **Data wrangle** static Web pages from different sources
- Explore a few of the mysteries of OAI-PMH
- Understand how to download web pages
- Wrap-up the boot camp



## A formal definition

*“The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Continued support of this work remains a cornerstone of the Open Archives program. Over time, however, the work of OAI has expanded to promote broad access to digital resources for eScholarship, eLearning, and eScience.”*


OAI Staff [1]

## Currently there are three projects[1]


- ResourceSync – a synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server's evolving resources
- Protocol for Metadata Harvesting (OAI-PMH) – The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that **expose structured metadata via OAI-PMH.**
- Object Reuse and Exchange (OAI-ORE) – Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video.

# From the “horse’s mouth”

the OAI Protocol for Metadata Harvesting



Van de Sompel, Herbert  
Los Alamos National Laboratory - Research Library



Load attached file.

# How does NASA implement OAI-PMH?

They have a few requests:

- Maximum of 100 records per request
- “Heavy” harvesting between 8PM-8AM EST
- At least 3 seconds between requests
- Use specific formats

Be polite.

The screenshot shows the NASA STI website page titled "Harvesting Data from NTRS". The page header includes "STI Scientific and Technical Information Program" and navigation links: HOME, TOOLS, STI AUDIENCES, ABOUT, CONTACT / HELP. The main heading is "Harvesting Data from NTRS". The text explains that NTRS provides dissemination of NASA STI to the widest audience possible by allowing NTRS information to be harvested by sites using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It also mentions that OAI-PMH defines a mechanism for information technology systems to exchange citation information using the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). NTRS is designed to accept and respond to automated requests using OAI-PMH. Automated requests only harvest citation information and not the full-text document images. It provides contact information for the STI Evaluation Desk for assistance. It also includes a section for "Use of Government Information" and a "NASA STI on YouTube" section with a video player.

[https://www.sti.nasa.gov/harvesting-data-from-ntrs/#.Wnpsdl\\_LDCI](https://www.sti.nasa.gov/harvesting-data-from-ntrs/#.Wnpsdl_LDCI)

# Same image.

The screenshot shows a web browser window displaying the NASA STI Program website. The page title is "Harvesting Data from NTRS". The URL in the address bar is "www.sti.nasa.gov/harvesting-data-from-nttrs/#.Wnpsdl\_LDCI". The page features a header with the STI logo and navigation links: HOME, TOOLS, STI AUDIENCES, ABOUT, CONTACT / HELP. The main content area is titled "Harvesting Data from NTRS" and contains the following text:

The NTRS promotes the dissemination of NASA STI to the widest audience possible by allowing NTRS information to be harvested by sites using the [Open Archives Initiative Protocol for Metadata Harvesting \(OAI-PMH\)](#). OAI-PMH defines a mechanism for information technology systems to exchange citation information using the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). NTRS is designed to accept and respond to automated requests using OAI-PMH. Automated requests only harvest citation information and not the full-text document images.

If you are interested in harvesting from NTRS and you have any comments or questions about the process, please contact the [STI Information Desk](#) for assistance.

If you harvest from our database, please cite the NASA STI Program as a source of data.

Sites interested in harvesting from NTRS should review the following guidance before harvesting:

**Use of Government Information**

The NTRS serves out unlimited, unclassified, publicly available NASA citations and full-text documents (PDFs). Persons, organizations, and sites interested in obtaining NASA information should review [Disclaimers](#), [Copyright Notice](#), [Terms and Conditions of Use](#) for assistance.

On the right side of the page, there are two social media widgets:

- NASA STI on Twitter:** Shows a tweet from @NASA\_STI: "Look up in February in honor of St. Valentine, and see mythical love stories, like Orion chasing the Pleiades across the heavens." It includes "Embed" and "View on Twitter" options.
- NASA STI on YouTube:** Shows a video thumbnail titled "ELaNs Educational Launch of ManosateLite 'Poly'" with the subtitle "Educational Launch of ManosateLite 'Launching Education into Space'".

`https://www.sti.nasa.gov/harvesting-data-from-nttrs/#.Wnpsdl_LDCI`



# Starting the protocol (1 of 2)

Using this URL:

`http://ntrs.nasa.gov/oai?verb=`

`ListIdentifiers&metadataPrefix=oai_dc`

Results in approx 700 lines that look like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
<responseDate>2018-02-07T03:16:58.230Z</responseDate>
<request verb="ListIdentifiers" metadataPrefix="oai_dc">http://ntrs.nasa.gov</request>
<ListIdentifiers>
<record>
<header>
<identifier>oai:casi.ntrs.nasa.gov:20180000827</identifier>
<datestamp>2018-01-31</datestamp>
<setSpec>All_STI</setSpec>
<setSpec>CASI_available_STI</setSpec>
</header>
</record>
<record>
<header>
<identifier>oai:casi.ntrs.nasa.gov:20180000826</identifier>
<datestamp>2018-01-31</datestamp>
<setSpec>All_STI</setSpec>
<setSpec>CASI_available_STI</setSpec>
</header>
```

## Starting the protocol (2 of 2)

```
...  
</OAI-PMH>
```

The important part is the identifier field value.

## With the identifier(s) (1 of 2)

Using the URL:

`http://ntrs.nasa.gov/oai?verb=GetRecord&identifier=20180000827&metadataPrefix=oai_tdc`  
Results in approximately 30 lines that look like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <responseDate>2018-02-07T03:23:12.426Z</responseDate>
  <request verb="GetRecord" identifier="20180000827" metadataPrefix="oai_dc">http://ntrs.nasa.gov</request>
  <GetRecord>
    <record>
      <header>
        <identifier>oai:casi.ntrs.nasa.gov:20180000827</identifier>
        <datestamp>2018-01-31</datestamp>
      </header>
      <metadata>
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
          <dc:identifier>Document ID: 20180000827</dc:identifier>
          <dc:title>Using Information from Rendezvous Missions For Best-Case Appraisals of Impact Damage t
          <dc:description>The Asteroid Threat Assessment Project (ATAP), a part of NASAs Planetary Defense
          <dc:date>20170515</dc:date>
          <dc:date>May 15, 2017</dc:date>
          <dc:rights>Copyright, Public use permitted</dc:rights>
          <dc:coverage>Unclassified, Unlimited, Publicly available</dc:coverage>
          <dc:identifier>http://hdl.handle.net/2060/20180000827</dc:identifier>
```

## With the identifier(s) (2 of 2)

```

<dc:source>CASI</dc:source>
<dc:format>application/pdf</dc:format>
<dc:creator>Burkhard, C. D.</dc:creator>
<dc:creator>Chodas, P. W.</dc:creator>
<dc:creator>Mathias, D. L.</dc:creator>
<dc:creator>Ulamec, S.</dc:creator>
<dc:creator>Arnold, J. O.</dc:creator>
<dc:type>ARC-E-DAA-TN42140-2</dc:type>
<dc:type>Annual IAA Planetary Defense Conference (PDC 2017); 15-19 May 2017; Tokyo; Japan</dc:type>
<dc:subject>Space Sciences (General)</dc:subject>
<dc:subject>Aeronautics (General)</dc:subject>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>

```

Where the important parts are the title and the description values.

## That we can have R pick apart into: (1 of 2)

...

```
$GetRecord$record$metadata
```

```
$GetRecord$record$metadata$dc
```

```
$GetRecord$record$metadata$dc$identifier
```

```
[1] "Document ID: 20180000827"
```

```
$GetRecord$record$metadata$dc$title
```

```
[1] "Using Information from Rendezvous Missions For Best-Ca
```

```
$GetRecord$record$metadata$dc$description
```

```
[1] "The Asteroid Threat Assessment Project (ATAP), a part
```

```
$GetRecord$record$metadata$dc$date
```

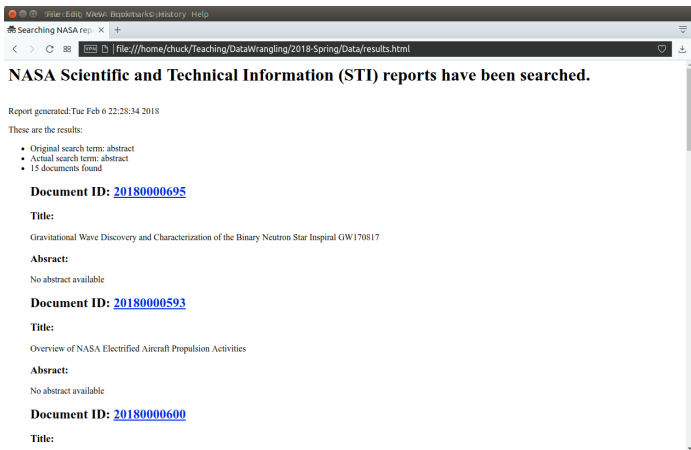
## That we can have R pick apart into: (2 of 2)

```
[1] "20170515"
```

```
...
```

We can work with the title and description.

# Which can lead to this:



Searching NASA rep: X +

file:///home/chuck/Teaching/DataWrangling/2018-Spring/Data/results.html

## NASA Scientific and Technical Information (STI) reports have been searched.

Report generated: Tue Feb 6 22:28:34 2018

These are the results:

- Original search term: abstract
- Actual search term: abstract
- 15 documents found

**Document ID:** [2018000695](#)

**Title:**

Gravitational Wave Discovery and Characterization of the Binary Neutron Star Inspiral GW170817

**Abstract:**

No abstract available

**Document ID:** [2018000593](#)

**Title:**

Overview of NASA Electrified Aircraft Propulsion Activities

**Abstract:**

No abstract available

**Document ID:** [2018000600](#)

**Title:**



# And ultimately to this:

The screenshot shows a web browser window displaying the NASA Technical Reports Server (NTRS) search results page. The browser's address bar shows the URL `nltrn | ntrs.nasa.gov/search.jsp`. The page features the NASA logo and social media icons (Facebook, Twitter, YouTube, RSS, and a plus sign). Below the header is a navigation menu with links for BASIC SEARCH, ADVANCED SEARCH, ABOUT NTRS, OAI HARVEST, SEARCH TIPS, and CONTACT / HELP.

The main content area includes a search bar with the text "Start a New Search: Enter search terms..." and a "Search" button. Below the search bar, the "Search Results" section displays the following information:

- Current Search Criteria:** Search: [Document ID]\*20180000695\* (with a "Remove All" link).
- Records:** A "Records" button and a "Concept Cloud" visualization.
- Matching Records:** 1
- Sort Results By:** A dropdown menu currently set to "Date Acquired". Other options include "1st Author", "Publication Date", and "Document ID".
- Text Size:** A button to adjust text size and a "Download to Computer" icon.
- Selected:** A "View Selected" link and a "Clear Selected" link.

The search results list one entry:

1. Gravitational Wave Discovery and Characterization of the Binary Neutron Star Inspiral GW170817  
 Document ID: [20180000695](#)  
 NTRS Full-Text: [Click to View](#) [PDF Size: 956 KB]  
 Author: Littenberg, Tyson B.  
 Abstract: No abstract available

On the right side of the page, there is a "Search History" panel showing the current search and a previous search for "[All]robot (9490)".



# Load the NASA reports script into RStudio

We'll do a few things with the script:

- 1 Modify the string tokenizer to be more robust
- 2 Modify the "searchTerm" variable to handle several terms at once
- 3 Modify the returnDocumentIDs() function to handle several terms at once
- 4 Discuss how to modify program to download 1,000,000 reports

## Q & A time.

Q: How many Harvard MBA's does it take to screw in a light bulb?

A: Just one. He grasps it firmly and the universe revolves around him.



## What have we covered?

- Explored a little bit of the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)
- Used R to download web pages
- Used R to extract data based on OAI-PMH tags



Next: Exploring the wild and woolly Web world.

## References (1 of 1)

- [1] OAI Staff, [Open Archives Initiative](https://www.openarchives.org),  
<https://www.openarchives.org>, 2018.

## Files of interest

1 Van de Sompel OAI PMH  
presentation 

2 NASA Reports 