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)

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

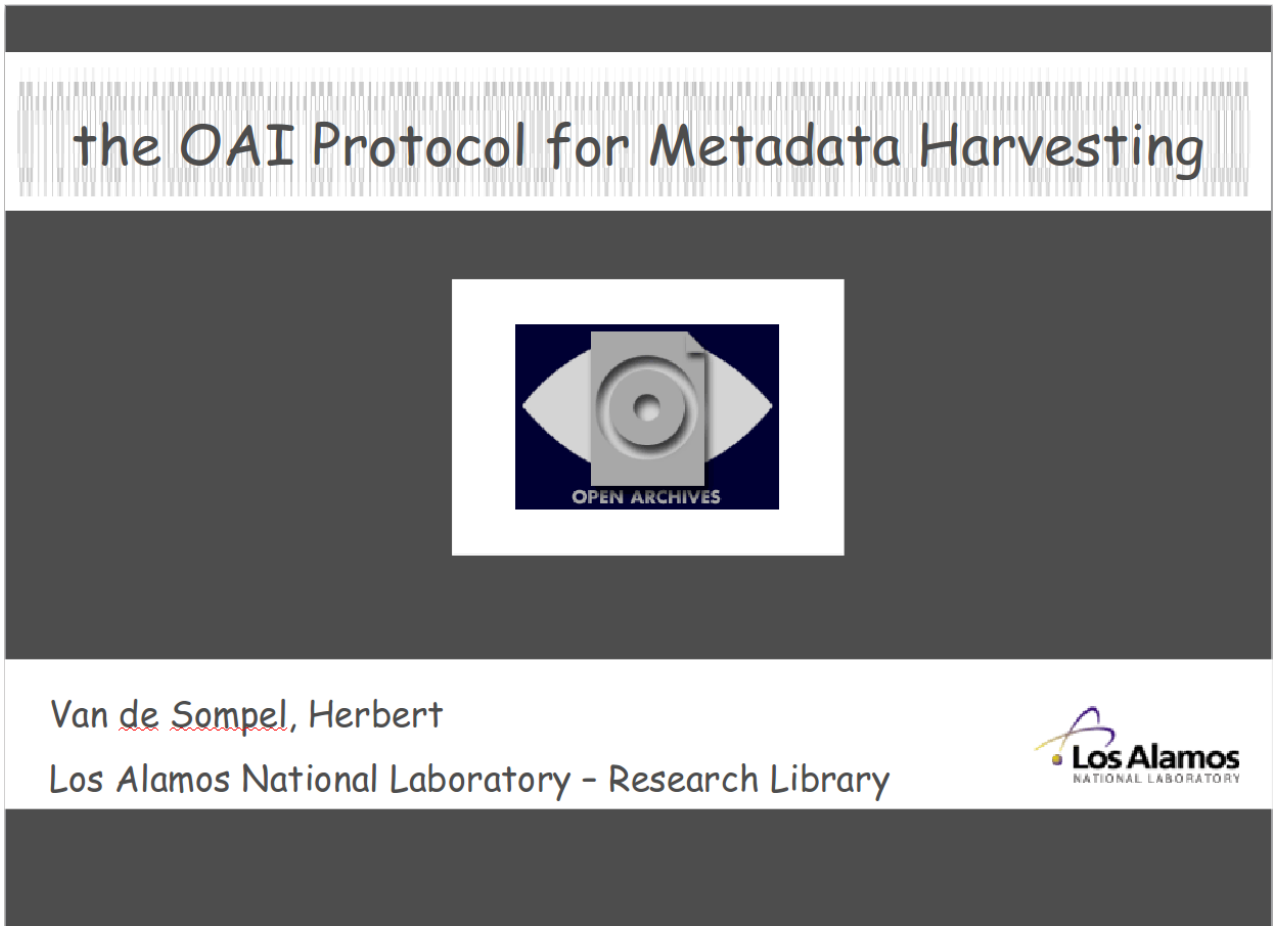
"The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. Continued support of this work remains a cornerstone of the Open Archives program. 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"



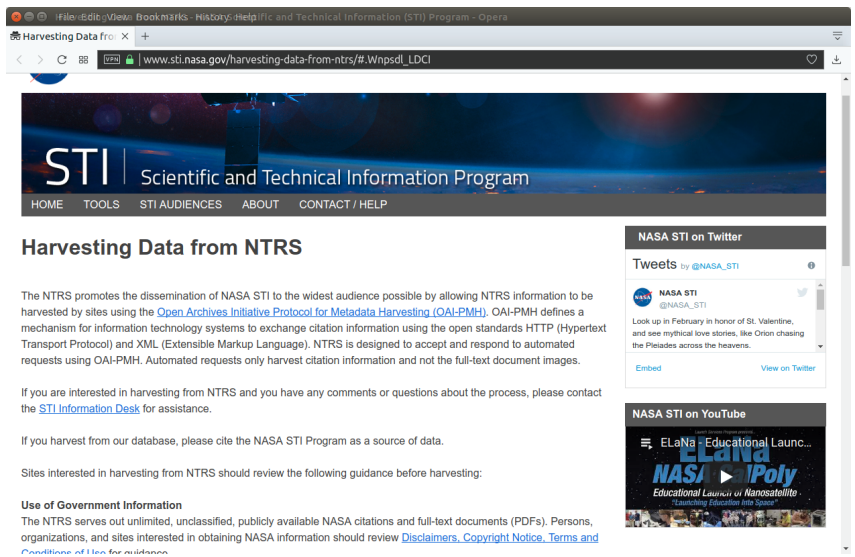
Load attached file.

How does NASA implement OAI-PMH?

They have a few requests:

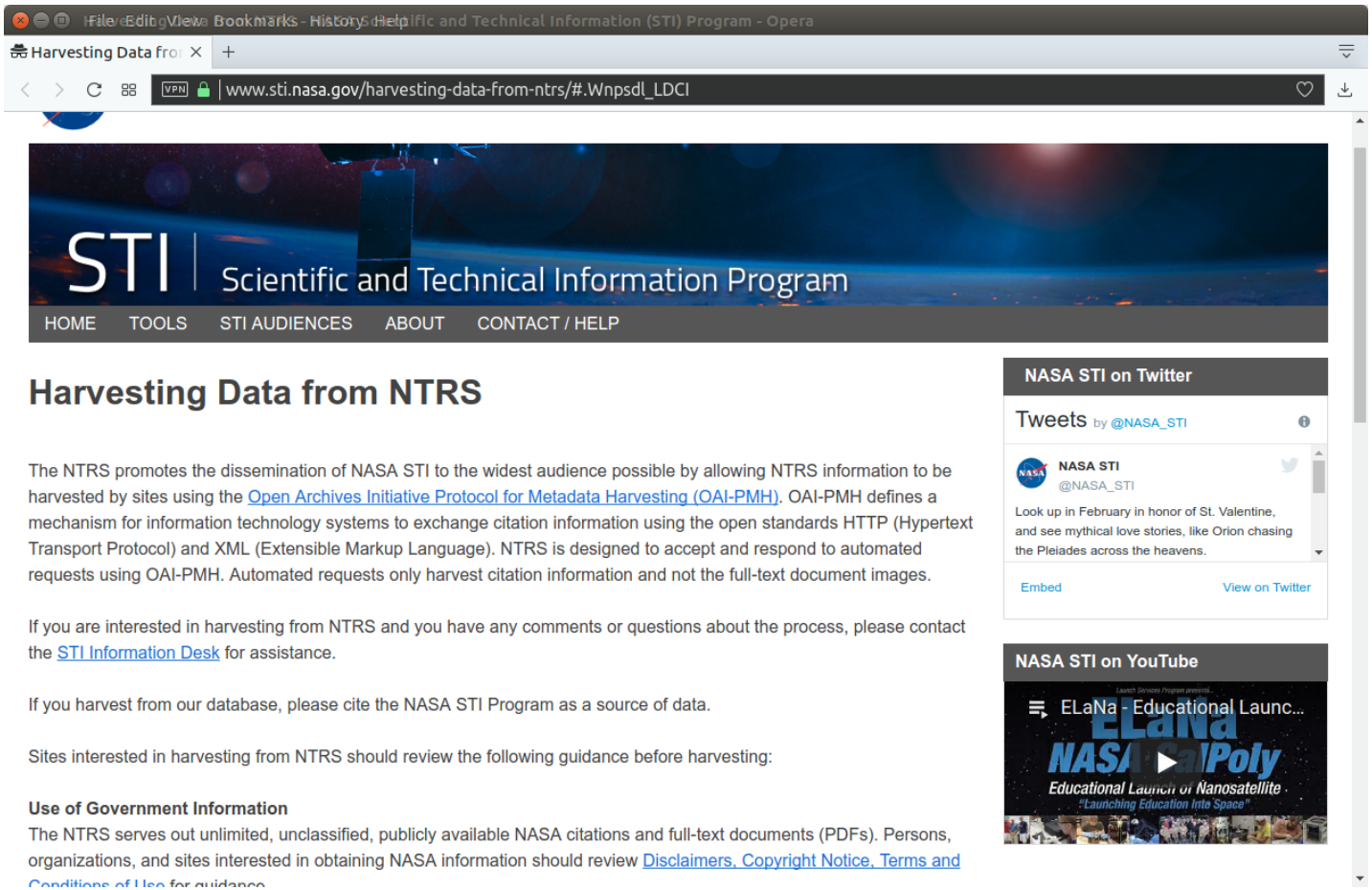
- Maximum of 100 records per request
- "Heavy" harvesting between 8PM-8AM EST
- At least 3 seconds between requests
- Use specific formats

Be polite.



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

Same image.



https://www.sti.nasa.gov/harvesting-data-from-ntrs/#.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</reques
  <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
          <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:ty
          <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)

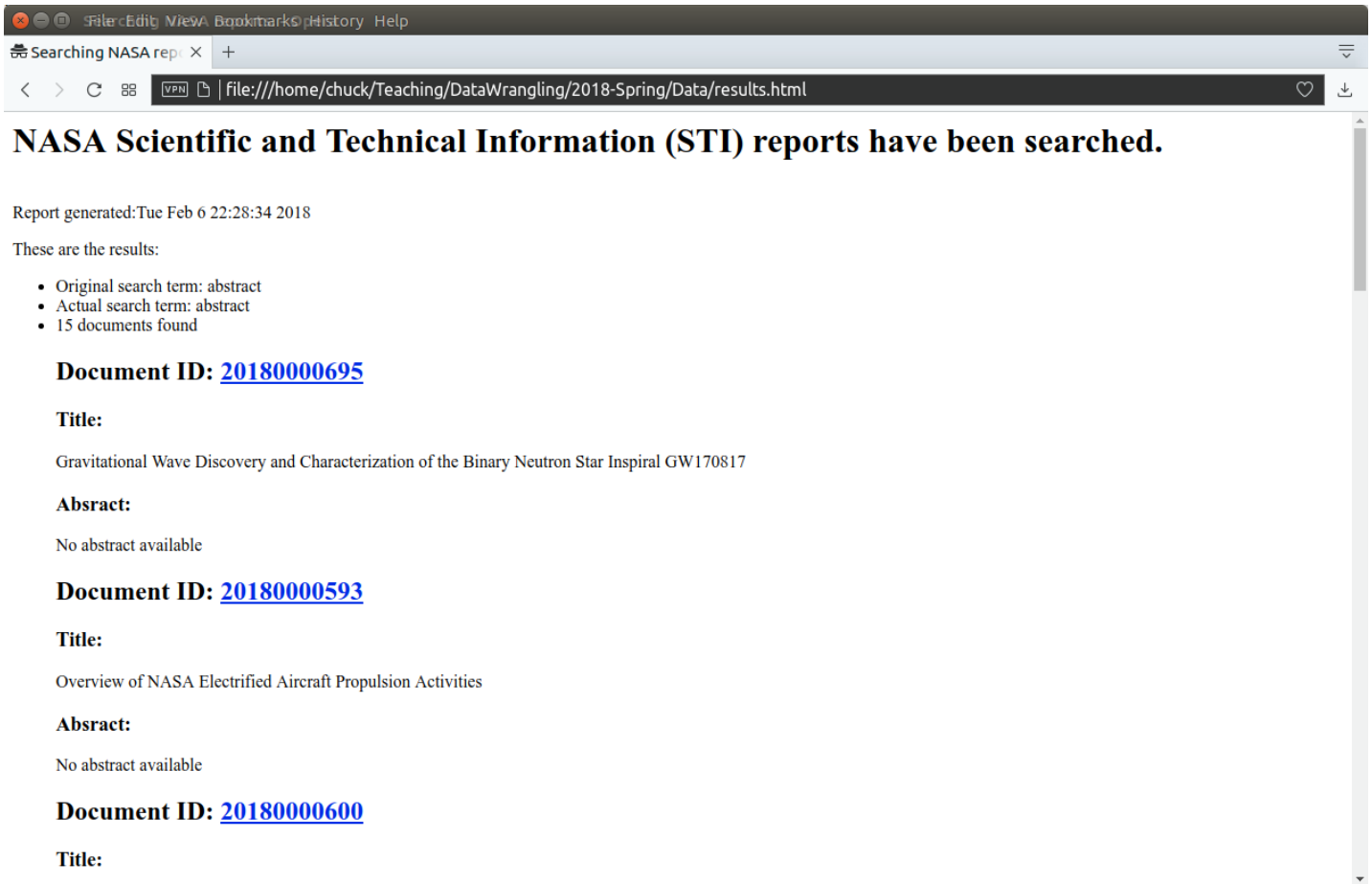
```
...
$GetRecord$record$metadata
$GetRecord$record$metadata$dc
$GetRecord$record$metadata$dc$identifier
[1] "Document ID: 20180000827"

$GetRecord$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)

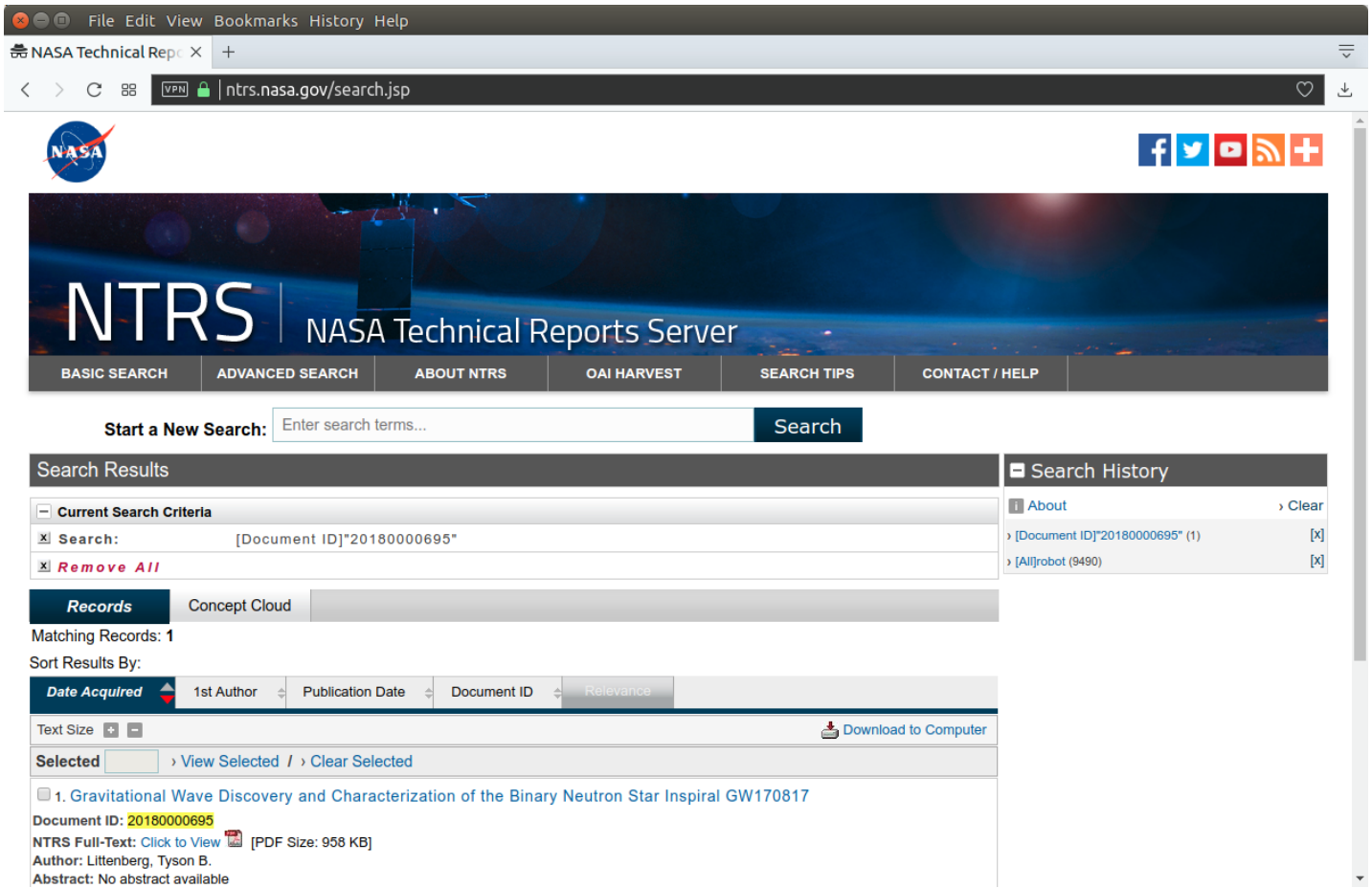
```
[1] "20170515"
...
```

We can work with the title and description.

Which can lead to this:



And ultimately to this:



Load the NASA reports script into RStudio

We'll do a few things with the script:

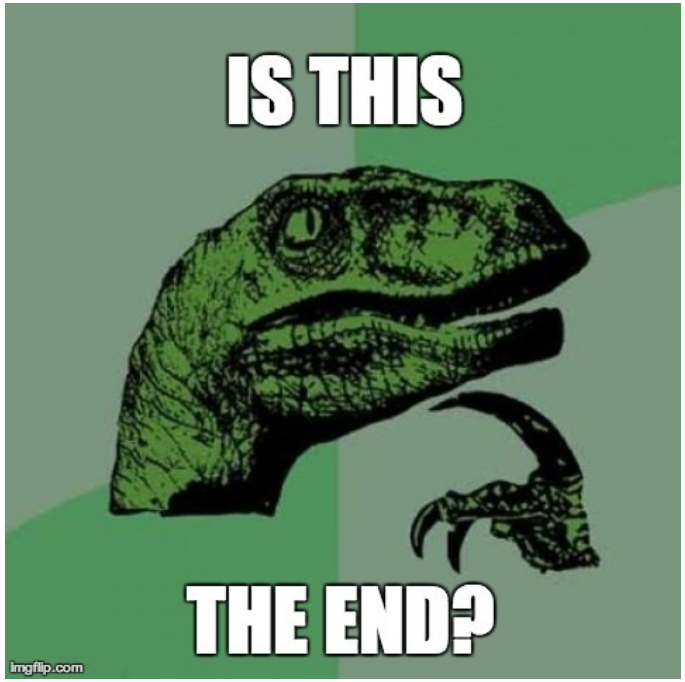
1. Modify the string tokenizer to be more robust
2. Modify the "searchTerm" variable to handle several terms at once
3. Modify the returnDocumentIDs() function to handle several terms at once
4. Discuss how to modify program to download 1,000,000 reports

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, 2018.

Files of interest

1. Van de Sompel OAI PMH presentation 
2. NASA Reports