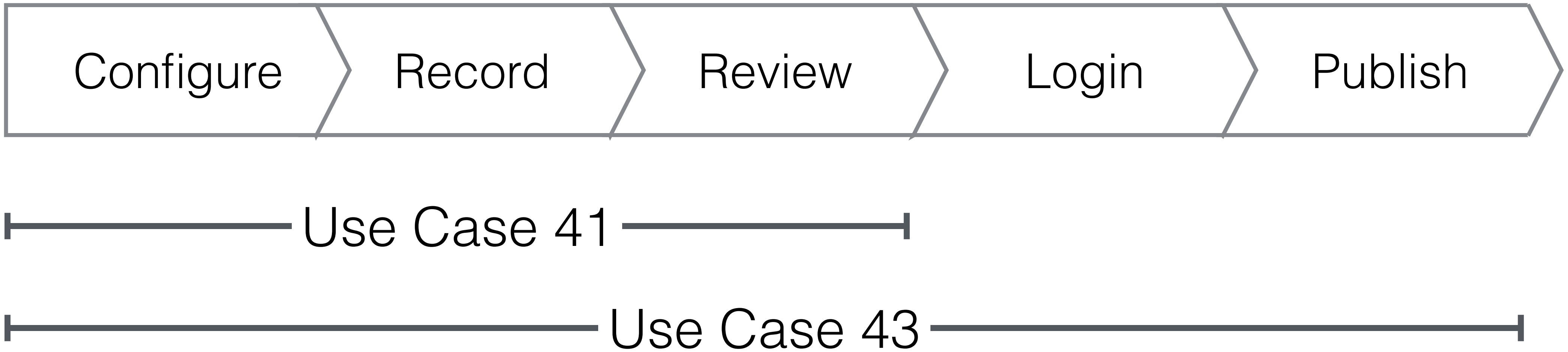
Provenance Scenarios

As a *<role>*, I want to *<goal>* so I can *<reason>*.

Use Cases 41, 43

As a *data analyst* using R or Matlab, I want to *keep track of my data input files, data output files and scripts* so I can *review my runs and potentially choose those to share with colleagues through an established DataONE repository*.

UC41, UC43



Configure

Record

Review

Login

Publish

Use Case 41

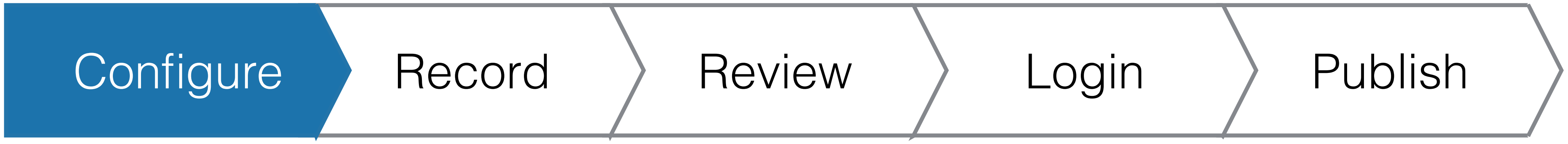
Use Case 43

Will be using Matlab code examples

Configure → Record → Review → Login → Publish

my_script.m

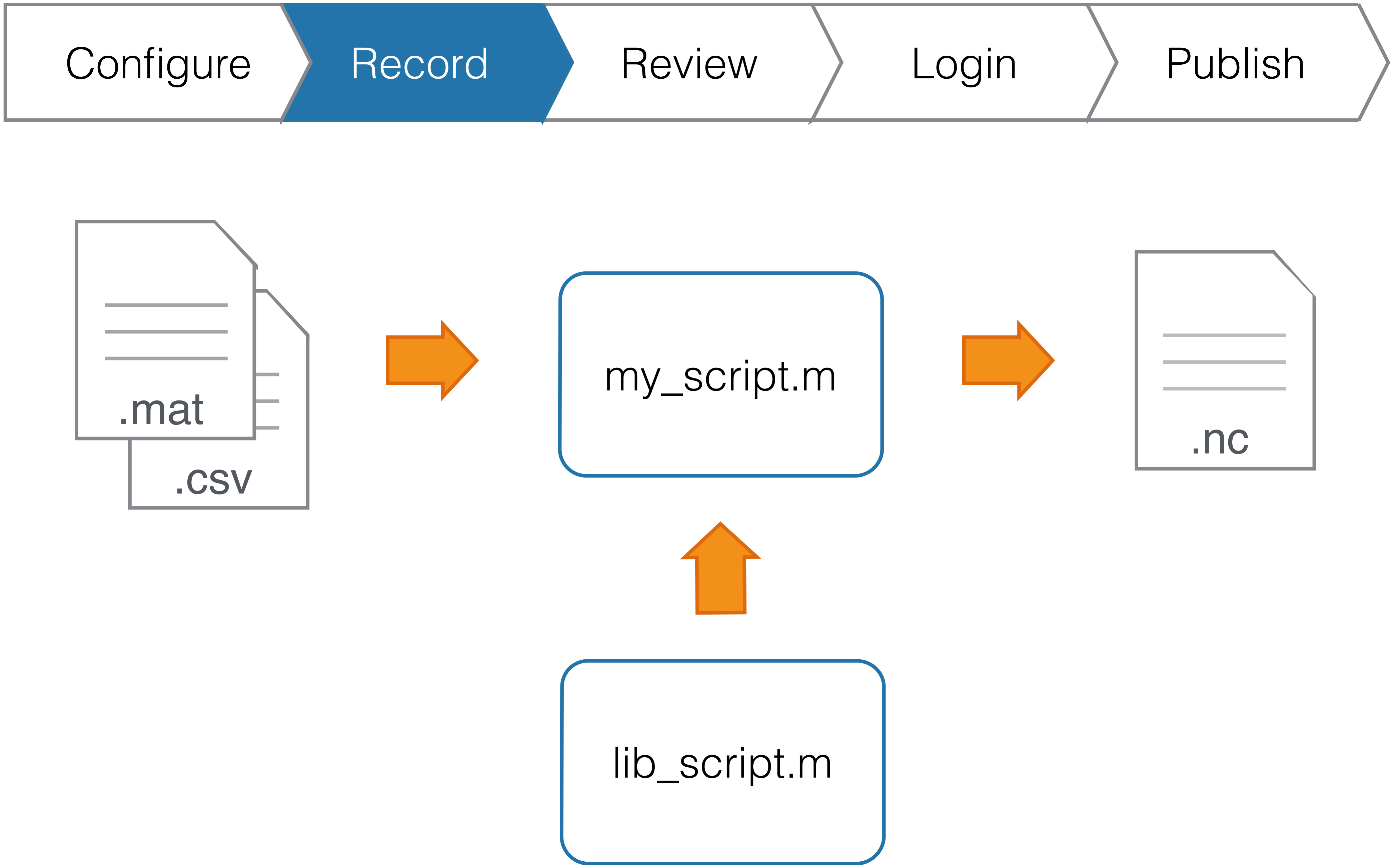
```
import org.dataone.client.configure.Configuration;

config = Configuration;

set(config, ...
     ‘baseDirectory’, ...
     ‘/Users/cjones/matlab/runs’);

set(config, ...
     ‘sourceRepositoryBaseURL’, ...
     ‘https://mercury-ops2.ornl.gov/MSTMIP/mn’);
```



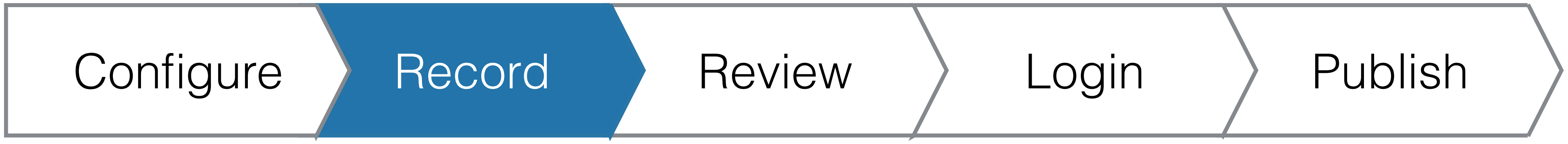
my_script.m

```
import org.dataone.client.configure.Configuration;
import org.dataone.client.configure.LabelParser;

parser = LabelParser;
parser.parse; % looks for config in comments

% d1prov:ingestStep
% prov:used, ‘/Users/cjones/data.nc’
some_array = ...
    get_input_data(‘/Users/cjones/data.nc’);
```

Alternative: develop a cross-language markdown-like annotation syntax and parsers



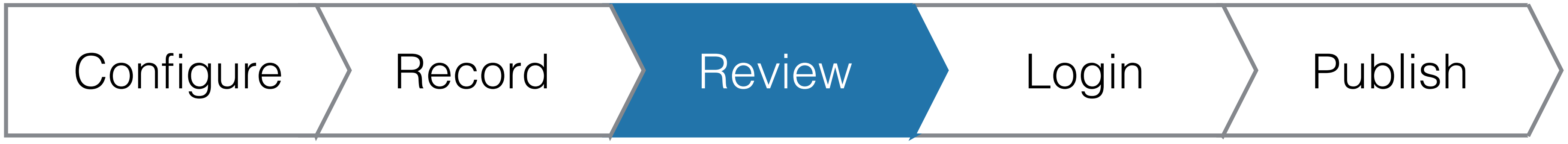
record() tracks input files, output files, referenced scripts, and the main script



my_script.m

```
import org.dataone.client.run.RunManager;

runManager = RunManager;

runId = runManager.record(‘/Users/cjones/my_script.m’);

% run your model analysis here
```

record() hides details of using insertRelationship() under the hood



Configure

Record

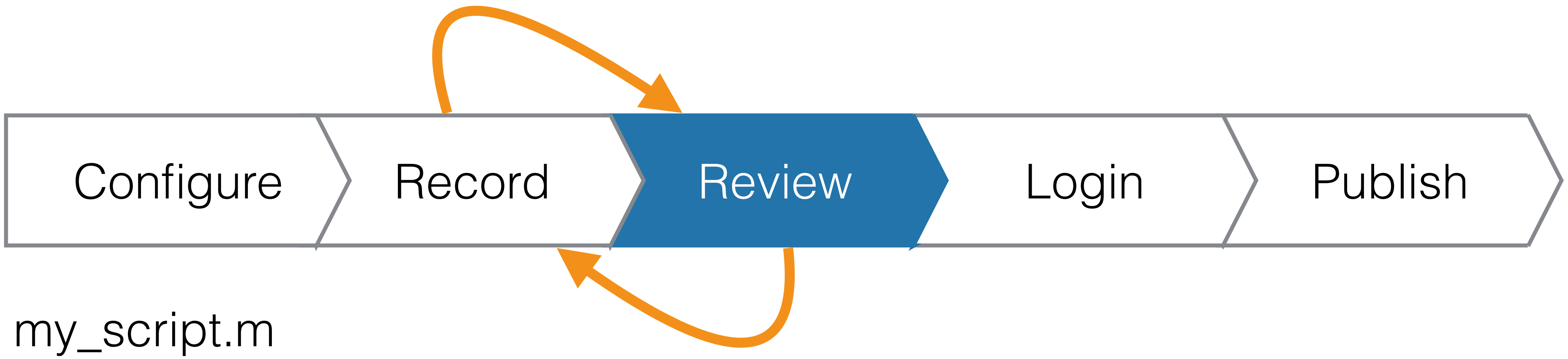
Review

Login

Publish

my_script.m

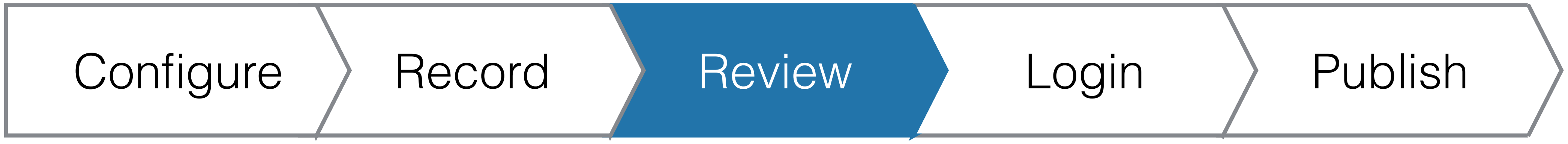
```
runManager.list(); % Prints out summary about all runs

runManager.view(runId); % More details about a run
```



my_script.m

```
runManager.list(); % Prints out summary about all runs

runManager.view(runId); % More details about a run
```

The record/review cycle is iterative

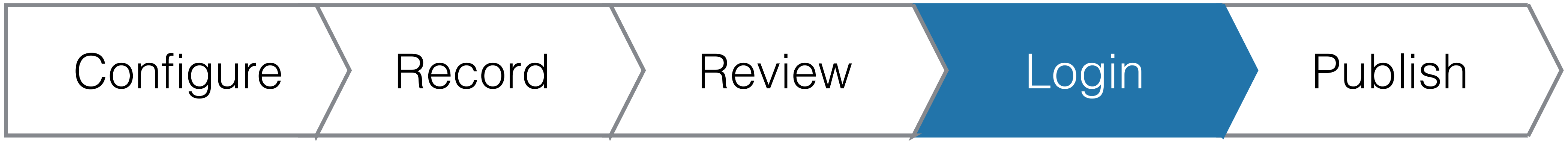


my_script.m

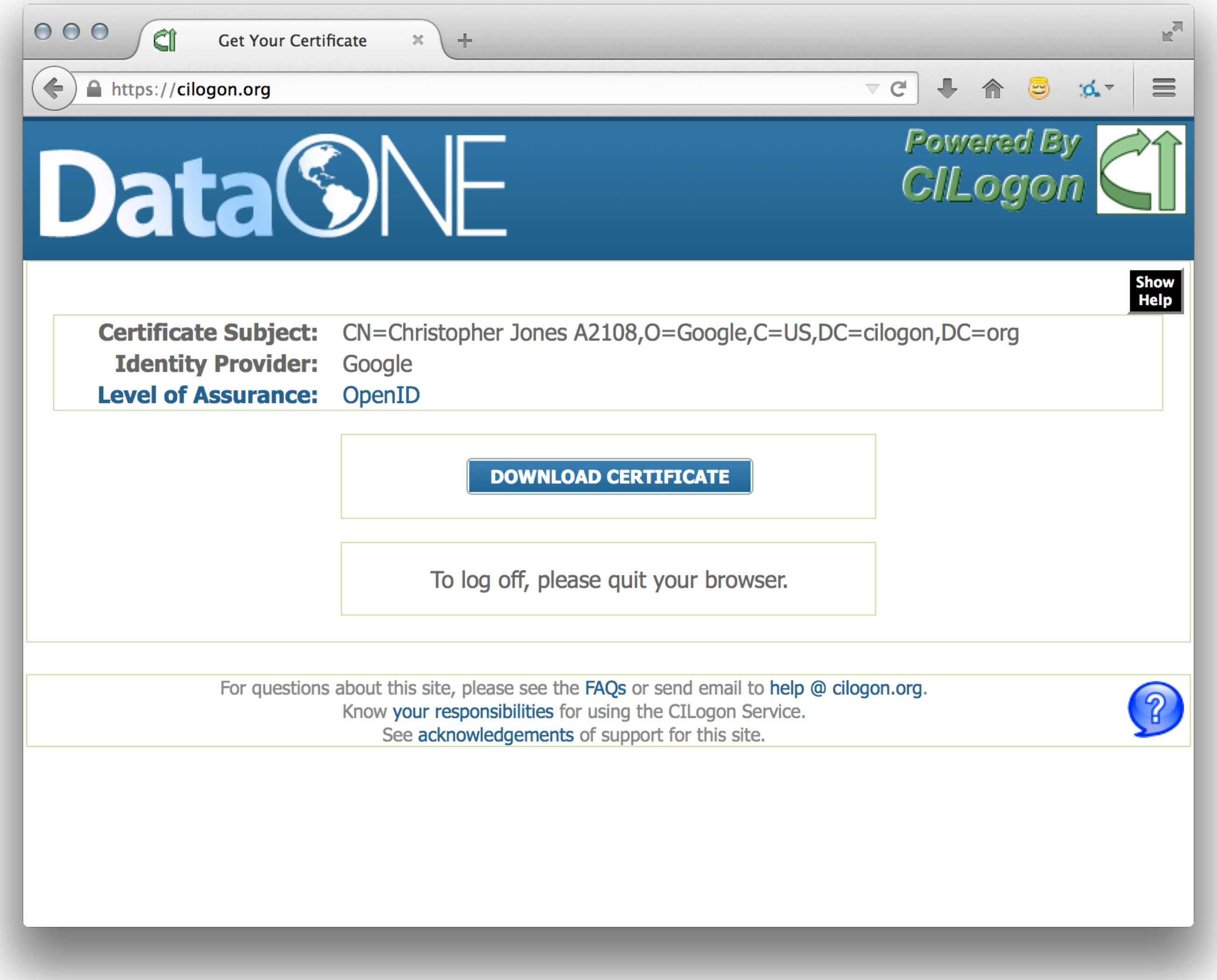
```
runManager.view(execution.1.1);
```

```
Run: execution.1.1
------------------
run start time:   Mon Sep 8 13:01
run end time:     Mon Sep 8 13:02
Matlab version:   2014a
Operating system: Mac OS X 10.9.5
Host name:        laurenshome

  Data Package: datapackage.1.1
  ------------------------------
  figure.1.1 was generated by execution.1.1
  execution.1.1 used data.1.1
  execution.1.1 used script.1.1
  data.1.1 is documented by metadata.1.1
  figure.1.1 is documented by metadata.1.1
  script.1.1 is documented by metadata.1.1
```

We need to to discuss what should be shown in summaries and details



Configure
Record
Review
Login
Publish



Certificate Subject: CN=Christopher Jones A2108,O=Google,C=US,DC=cilogon,DC=org
Identity Provider: Google
Level of Assurance: OpenID

To log off, please quit your browser.

For questions about this site, please see the FAQs or send email to help @ cilogon.org.
Know your responsibilities for using the CILogon Service.
See acknowledgements of support for this site.

Login to DataONE and download your certificate

Configure | Record | Review | Login | Publish

my_script.m

```
% Publish the data package to the repository
runManager.publish(‘datapackage.1.1’);

```

RunManager will:

- create identifiers for each Data Package member
- upload each to the target DataONE Member Node
- upload the Data Package with all PROV relationships

UC 41/43 Todos

- Discuss the envisioned process (12-month goals)
  - Will it work? Are there gaps?
- Configuration Step:
  - Decide on programmatic vs inline-comment
  - Determine what configuration options are required
- Review Step
  - Discuss what is shown in summary and detail views

Use Case 42

As a *data analyst* using R or Matlab, *I want to* easily document my script with standardized comments *so I can* understand the workflow of the script.

UC42

Use Case 43

As a data analyst using R or Matlab, I want to publish my data and their history so I can share them with colleagues through an established DataONE repository.

UC43

Use Case 44

As a scientist, I want to be able to examine the original datasets used in a derived dataset I’ve found through DataONE so I can understand the history and composition of the derived dataset.

UC44

Use Case 45

As a scientist, I want to be able to find all derived datasets in DataONE that use my dataset so I can understand how my data are being used and by which colleagues.

UC44

Use Case 46

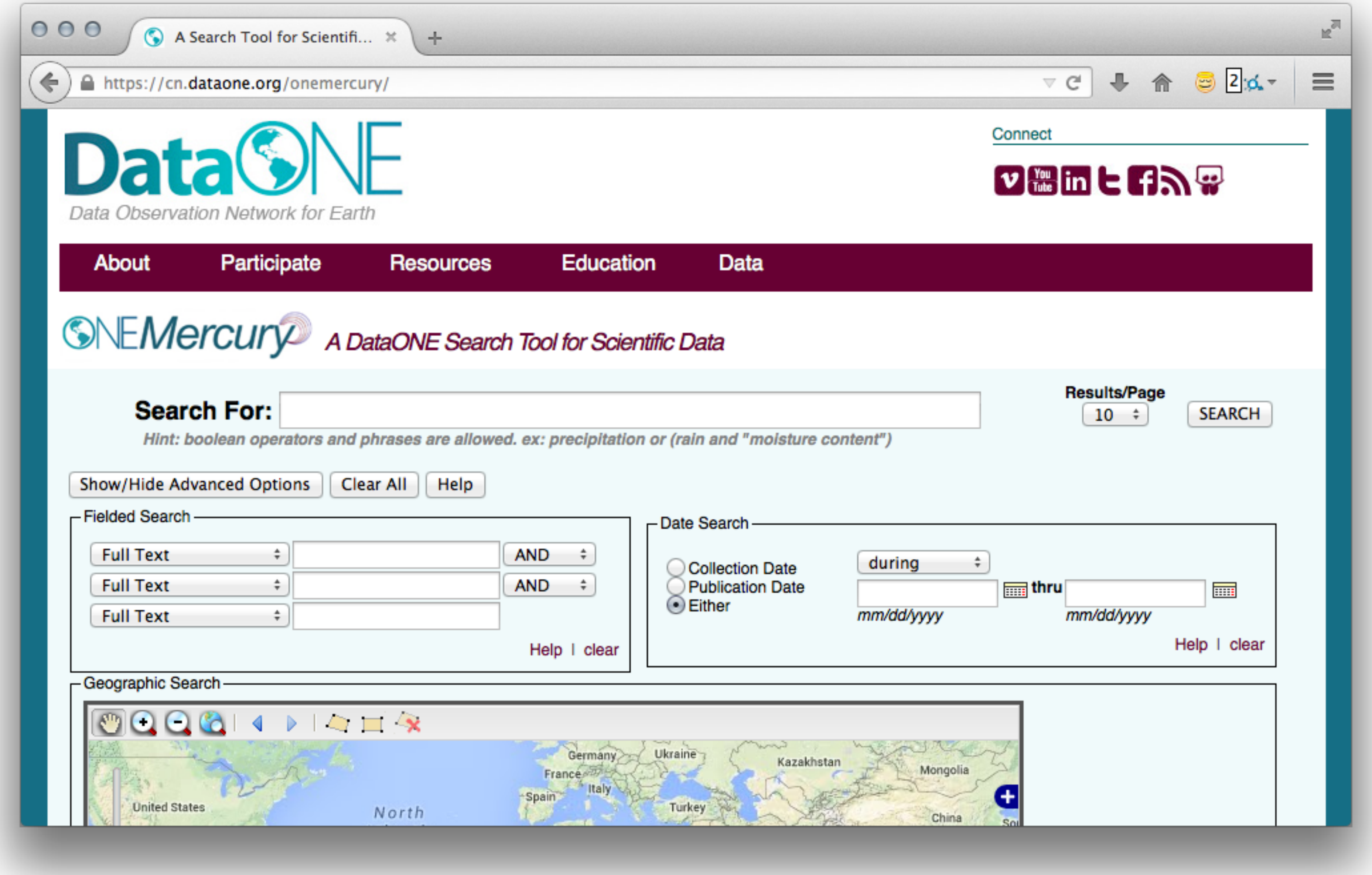
As a scientist reviewing derived tables or figures, I want to be able to examine the original datasets and the original script used to generate them so I can understand their history and composition.

UC44

Search

View

Download


A Search Tool for Scientifi...
https://cn.dataone.org/onemercury/
DataONE
Data Observation Network for Earth
Connect
About
Participate
Resources
Education
Data

ONEMercury *A DataONE Search Tool for Scientific Data*

Search For:

Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and "moisture content")

Results/Page: 10

Fielded Search

Full Text		AND
Full Text		AND
Full Text		

Date Search

○ Collection Date
○ Publication Date
● Either

during

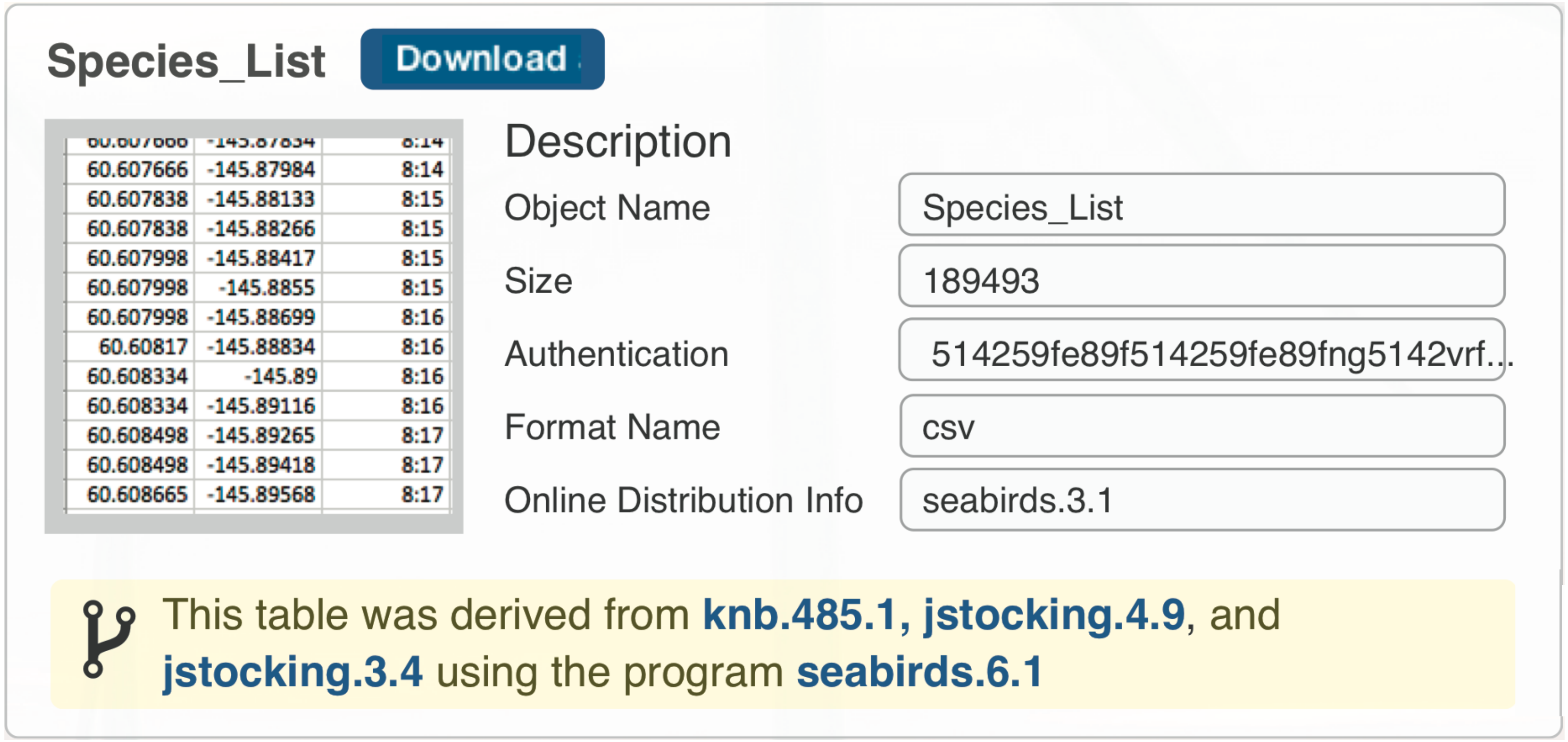
thru

mm/dd/yyyy

mm/dd/yyyy

Search
View
Download

Doe, John. 2014. Searbirds of the Gulf of Alaska and North Pacific. (seabirds.2.1)



Species_List Download

60.607666	-145.87834	8:14
60.607666	-145.87984	8:14
60.607838	-145.88133	8:15
60.607838	-145.88266	8:15
60.607998	-145.88417	8:15
60.607998	-145.8855	8:15
60.607998	-145.88699	8:16
60.60817	-145.88834	8:16
60.608334	-145.89	8:16
60.608334	-145.89116	8:16
60.608498	-145.89265	8:17
60.608498	-145.89418	8:17
60.608665	-145.89568	8:17

Description

Object Name	Species_List
Size	189493
Authentication	514259fe89f514259fe89fng5142vrf...
Format Name	csv
Online Distribution Info	seabirds.3.1

This table was derived from **knb.485.1, jstocking.4.9**, and **jstocking.3.4** using the program **seabirds.6.1**

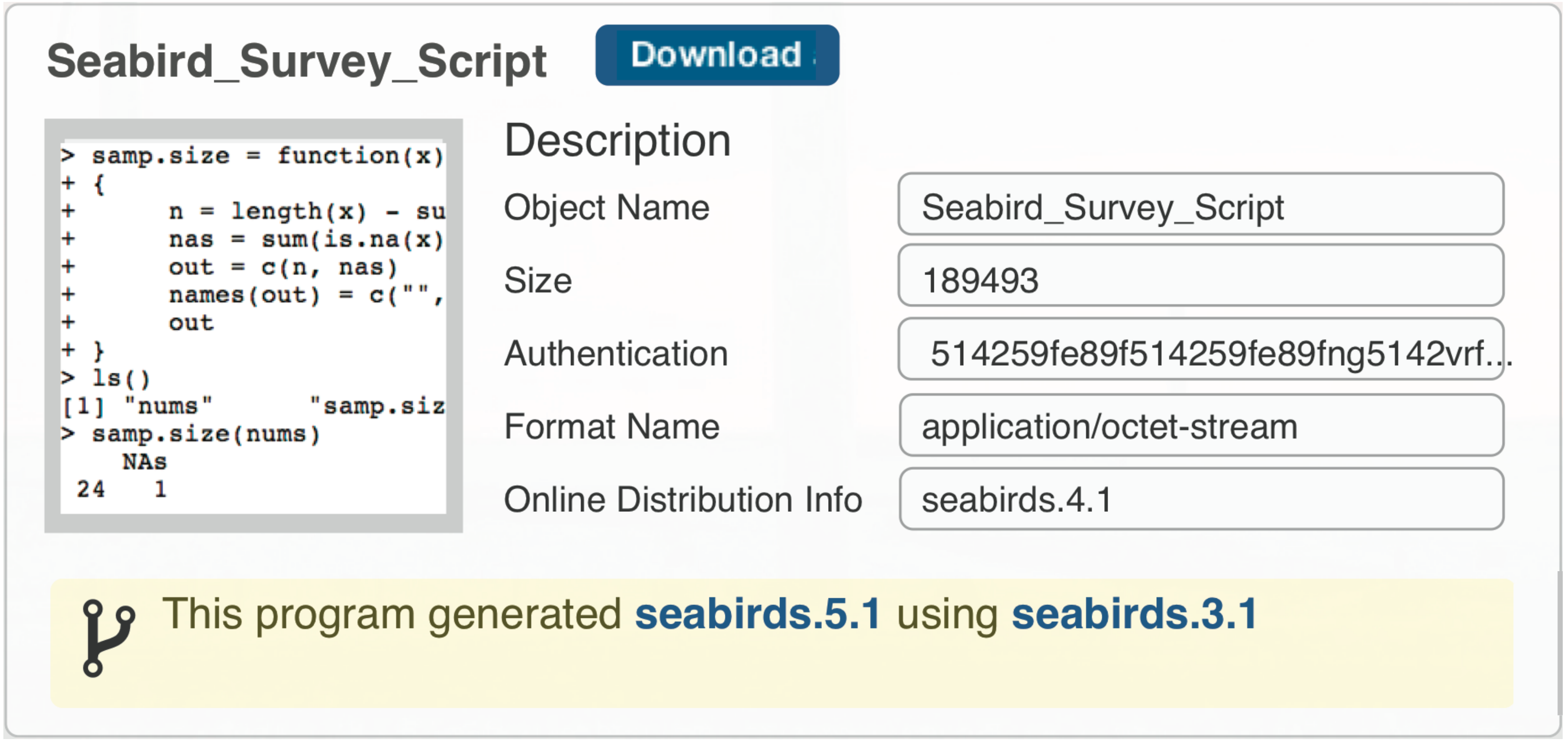
View derivation history

Search

View

Download

Doe, John. 2014. Searbirds of the Gulf of Alaska and North Pacific. (seabirds.2.1)



Seabird_Survey_Script

```
> samp.size = function(x)
+ {
+         n = length(x) - su
+         nas = sum(is.na(x)
+         out = c(n, nas)
+         names(out) = c("",
+         out
+ }
> ls()
[1] "nums"       "samp.siz
> samp.size(nums)
    NAs
 24    1
```

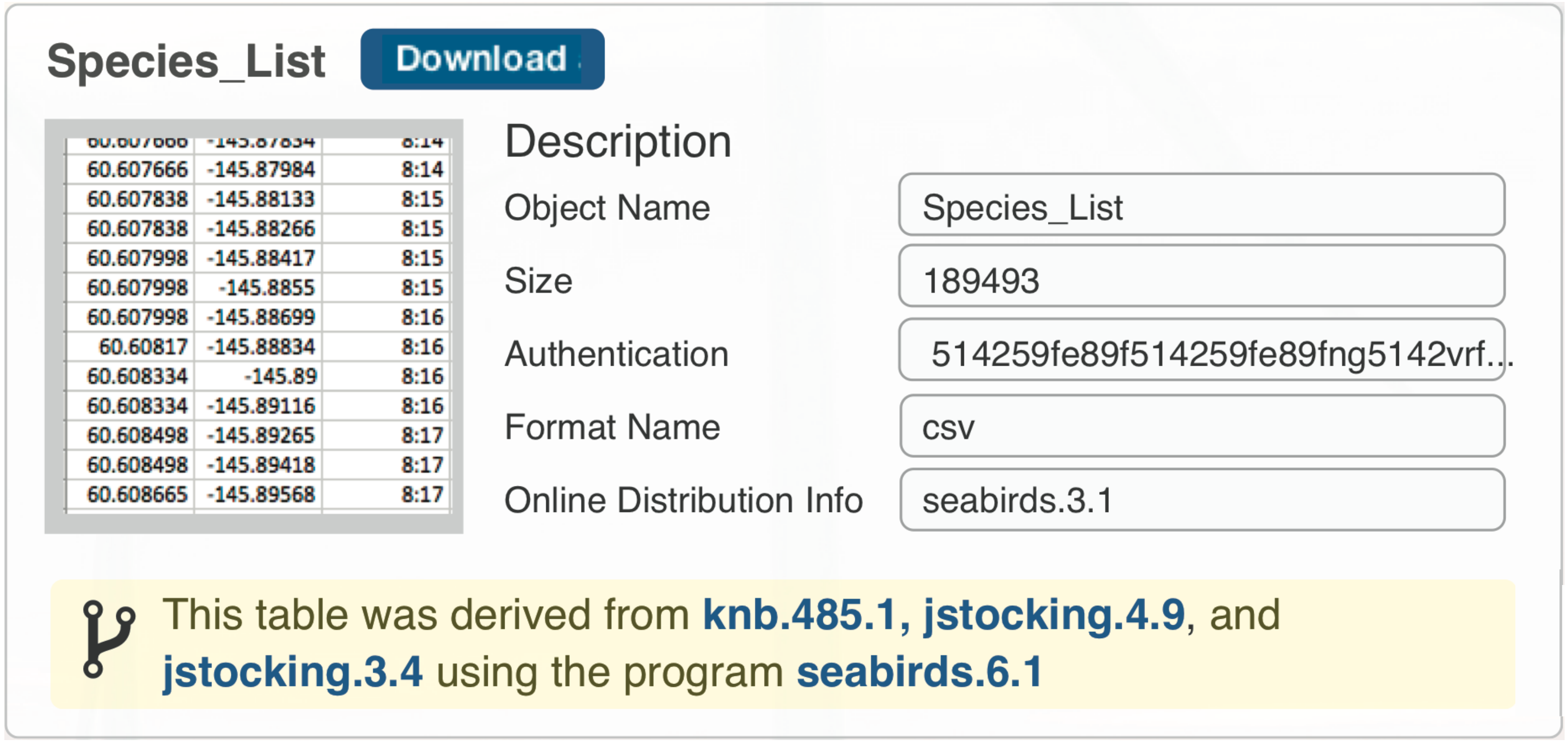
Description

Object Name	Seabird_Survey_Script
Size	189493
Authentication	514259fe89f514259fe89fng5142vrf...
Format Name	application/octet-stream
Online Distribution Info	seabirds.4.1

This program generated **seabirds.5.1** using **seabirds.3.1**

View generation history

Doe, John. 2014. Searbirds of the Gulf of Alaska and North Pacific. (seabirds.2.1)



Species_List

60.607666	-145.87834	8:14
60.607666	-145.87984	8:14
60.607838	-145.88133	8:15
60.607838	-145.88266	8:15
60.607998	-145.88417	8:15
60.607998	-145.8855	8:15
60.607998	-145.88699	8:16
60.60817	-145.88834	8:16
60.608334	-145.89	8:16
60.608334	-145.89116	8:16
60.608498	-145.89265	8:17
60.608498	-145.89418	8:17
60.608665	-145.89568	8:17

Description

Object Name	Species_List
Size	189493
Authentication	514259fe89f514259fe89fng5142vrf...
Format Name	csv
Online Distribution Info	seabirds.3.1

This table was derived from **knb.485.1**, **jstocking.4.9**, and **jstocking.3.4** using the program **seabirds.6.1**

Follow links to download

UC 42/43/44 Todos

- Provenance UI Design
  - Review and Modify (tomorrow morning session)