Reconstituting a Digital Repository through Use Case Driven Ontology Development

SIG-CM Workshop
JCDL 2019 June 6th, 2019

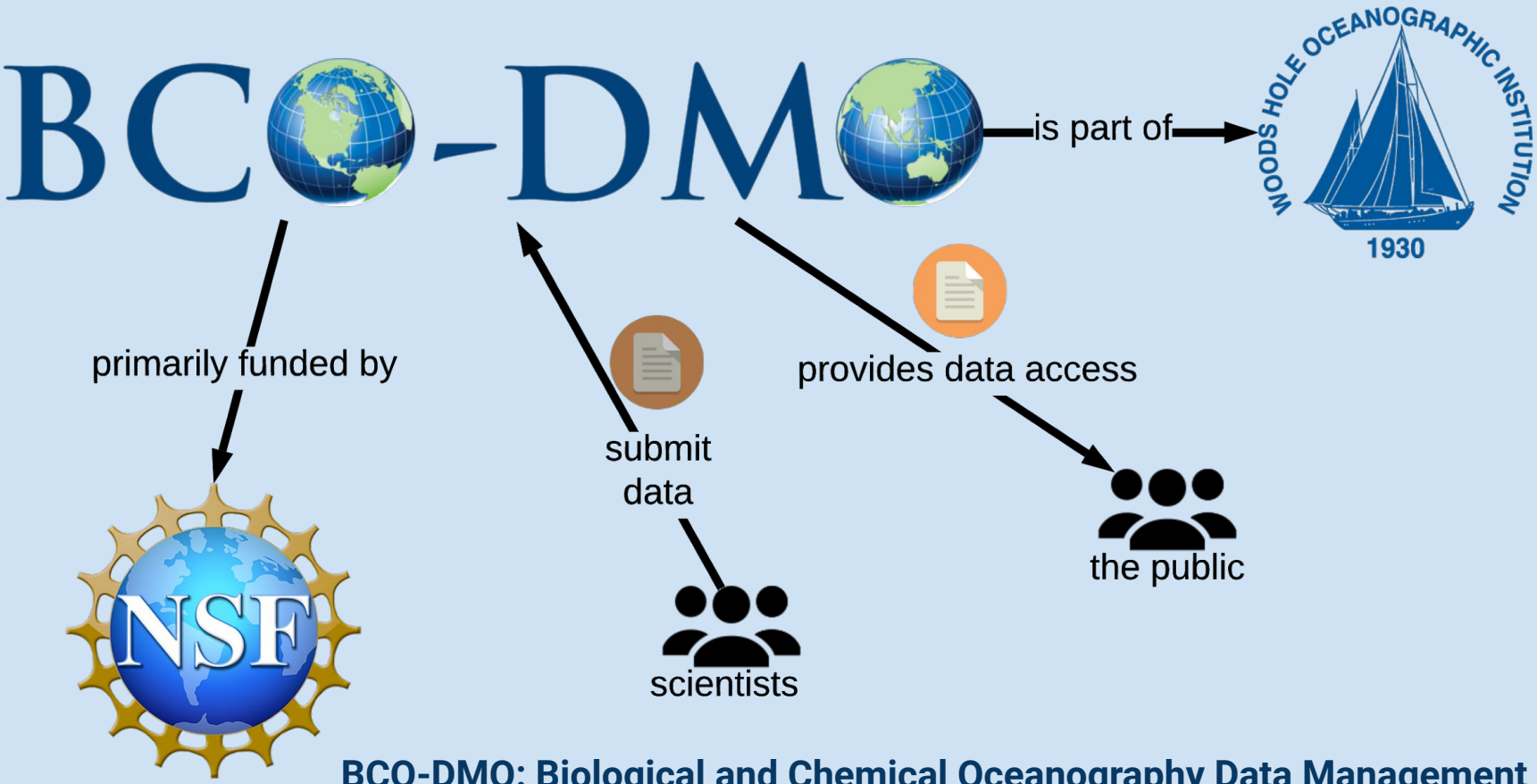
Adam Shepherd, Danie Kinkade, Douglas Fils

BCO-DMO

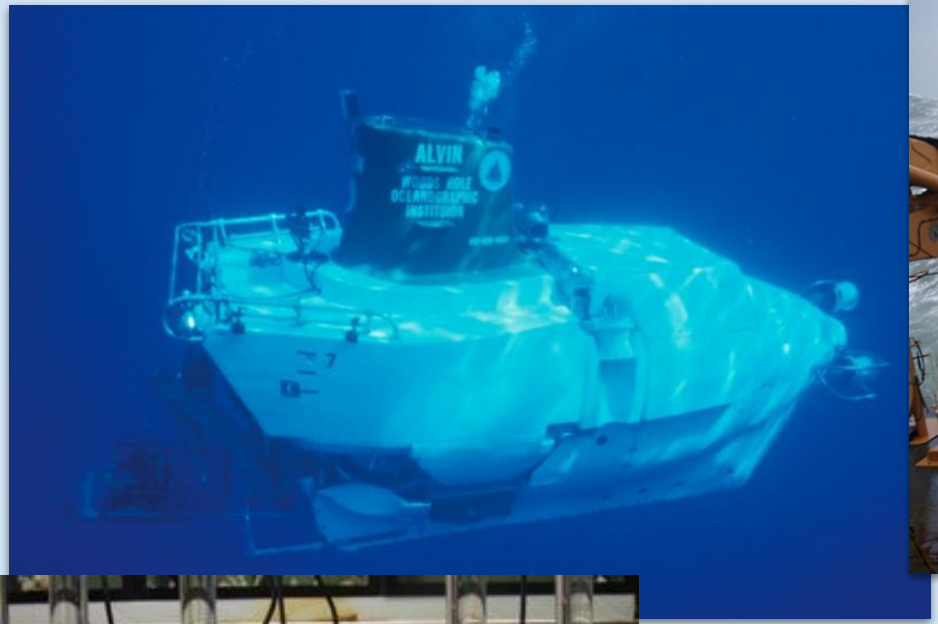
NSF
OCE-1435578

Woods Hole Oceanographic Institution

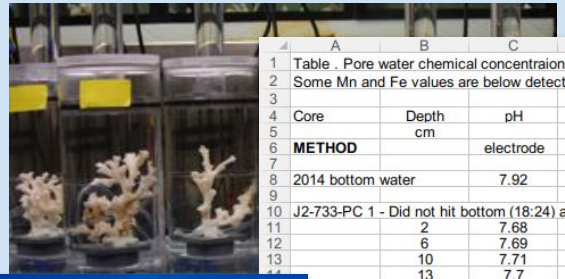
Who are we?



BCO-DMO: Biological and Chemical Oceanography Data Management Office


ALVIN
WOODS HOLE
OCEANOGRAPHIC
INSTITUTION


Did you record the
metadata?



	A	B	C	D
1	Table . Pore water chemical concentraion data and lo			
2	Some Mn and Fe values are below detection and are			
3				
4	Core	Depth	pH	alkalinity
5		cm		mmol/kg
6	**METHOD**		electrode	titration
7				
8	2014 bottom water		7.92	2.32
9				
10	J2-733-PC 1 - Did not hit bottom (18:24) and was pos			
11		2	7.68	2.08
12		6	7.69	2.11
13		10	7.71	2.20
		13	7.7	2.22
		16	7.69	2.22
		18	7.71	2.22
	-733-PC 2 - Did not hit bottom (18:27) and was positioned next to PC 1, 22 d			
		2	7.69	2.13	21.9	548.5
		7	7.70			
		11	7.73			
		15	7.70			
		18	7.72			
		20	7.70			
		22	7.72			
	-733-PC 4 - Hit bottom (18:01) and wa			
		3	7.67			
29		6	7.69			
30		8	7.71			
			7.75			
			7.71			
			7.73			
			it bottom (19:28)			
			7.71			
			7.71			

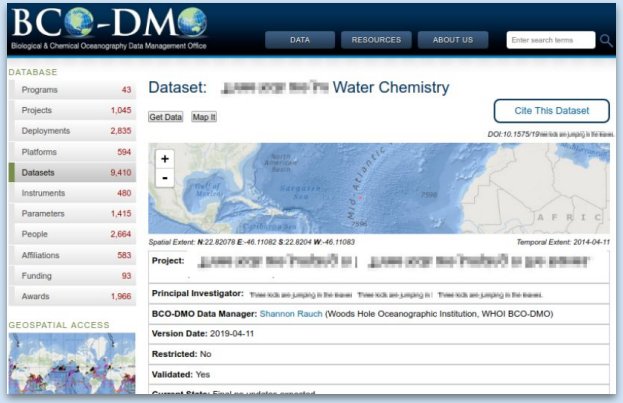
	A	B	C
1	Site Code	Site Code	Deployment Dates
2	1	Dittlif Point	6/1/16 - 3/22/17
3			3/27/17 - 6/22/17
4	2	Cocoloba Cay	5/29/16 - 3/22/17
5			3/27/17 - 7/11/17
6	3	Joel's Shoal	5/29/16 - 10/22/16
7			11/10/16 - 3/22/17
8			3/28/16 - 7/11/17
9	4	White Point	5/29/16 - 10/21/16
10			10/23/16 - 3/23/17
11	5	Europa Bay	5/29/16 - 10/21/16
12			10/23/16 - 12/12/16
13	6	Tektite	5/29/16 - 10/21/16
14			11/5/16 - 3/20/17
15			3/28/17 - 6/14/17
16			
17	7	Ya	
18			
19	8	Bo	
20			
21			
22	9	Ra	
23			
24			
25	S	Re	
26			
27			

C	D	E	F	G	H
Video ID	**Time (minute)**	**E**	**D**	**C.T.**	**C.O.**
Cnidaria species:	Moon Jellyfish (Aurelia aurita)				
Individual:	1				
Video ID	**Time (minute)**	**E**	**D**	**C.T.**	**C.O.**
Clip015	4.06			X	
	13.04	X			
	26.23				X
	30.17			X	
	45.09			X	
	1.34.06				X
Clip016	10.06				X
	27.45				X
	32.22	X			
	50.13				X
Clip017	10.06				X
	27.2				
	25.45	X			
	30.24				
	36.02				
	39.15	XX			
	43.19				
	1.06.27				

data.csv

```
Core,Depth,pH,alkalinity,Nitrate,Chlorinity,Ca,B,
2014 bottom water,,7.92,2.32,21.1,544.9,10.17,413
J2-733-PC 1,2,7.68,2.08,22.3,546.2,9.69,524,<0.1,
J2-733-PC 1,6,7.69,2.11,23.8,546.2,9.64,535,0.4,<
J2-733-PC 1,10,7.71,2.2,25.1,545.2,9.59,533,<0.1,
J2-733-PC 1,13,7.7,2.22,25.8,547.2,9.62,531,0.2,<
J2-733-PC 1,16,7.69,2.22,24.7,544.6,9.67,529,0.5,
J2-733-PC 1,18,7.71,2.22,24.6,546.6,9.67,525,0.2,
J2-733-PC 2,2,7.69,2.13,21.9,548.5,9.72,528,0.3,<
J2-733-PC 2,7,7.7,2.17,24.6,543.9,9.65,536,<0.1,<
J2-733-PC 2,11,7.73,2.18,25.5,546.2,9.62,532,<0.1
J2-733-PC 2,15,7.7,2.16,26.1,544.2,9.6,530,0.3,<0
J2-733-PC 2,18,7.72,2.14,25.7,545.9,9.64,519,0.4,
J2-733-PC 2,20,7.7,2.16,25.4,546.3,9.62,527,0.2,<
J2-733-PC 2,22,7.72,2.16,25.2,,9.63,525,<0.1,<0.1
J2-733-PC 4,3,7.67,2.05,23,547.3,9.7,521,<0.1,<0.
J2-733-PC 4,6,7.69,2.1,23.8,545.1,,516,0.2,<0.1,9
J2-733-PC 4,8,7.71,2.11,24.5,544,9.66,516,0.2,<0.
J2-733-PC 4,10,7.75,2.13,25.1,544.2,9.64,517,0.2,
J2-733-PC 4,12,7.71,2.13,25.3,544,9.61,514,0.1,<0
```

Dataset landing page



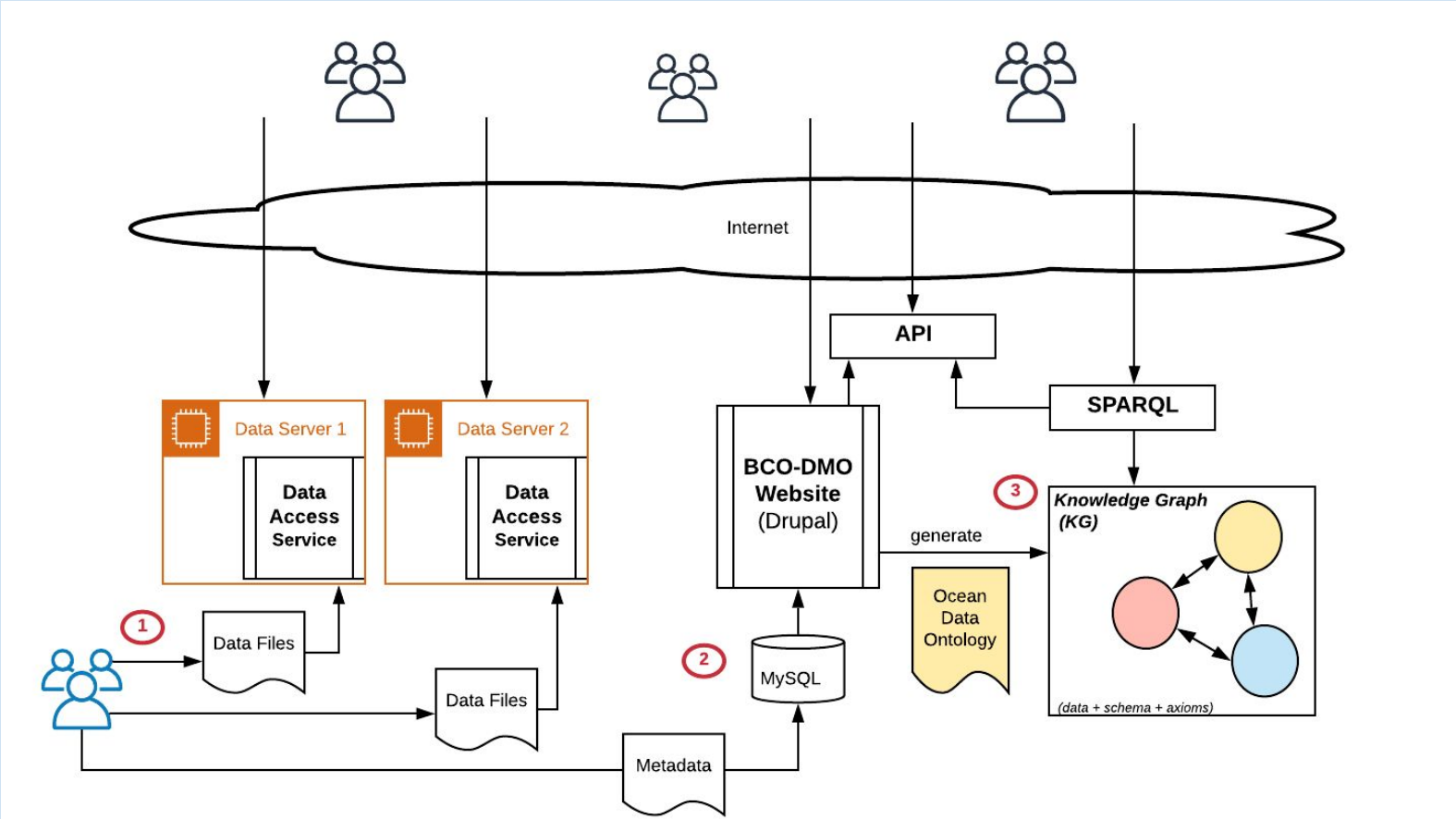
Why we needed a new infrastructure?

- **Big Data challenges**
  - Some data may be stored in other other external repositories
  - But, our infrastructure depends on local management
- **Software ages faster than data**

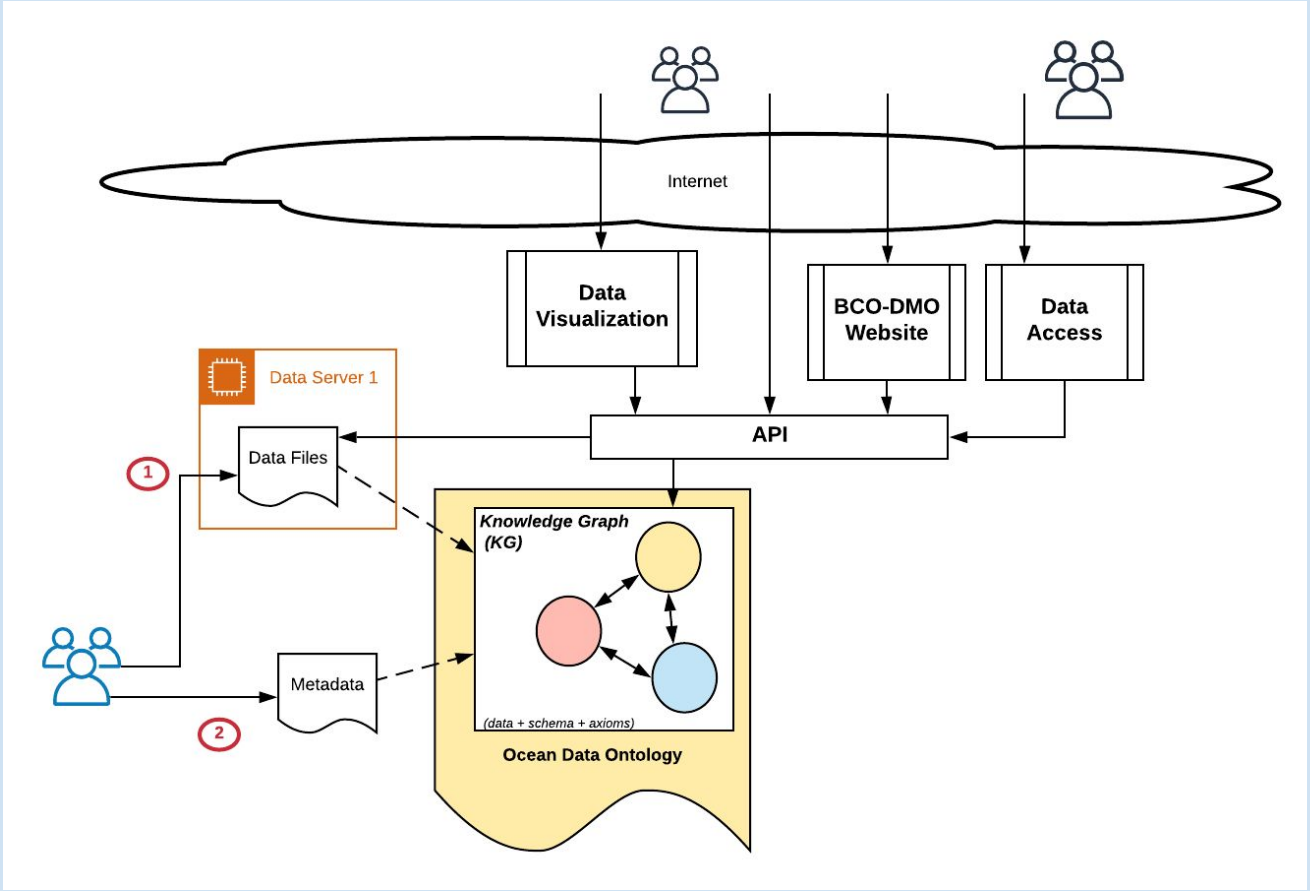
Data Submission
Metadata Catalog
Data Access
Data Visualization

- nearing end-of-life
- dependent on each other

OLD: Flat Architecture = Dependency

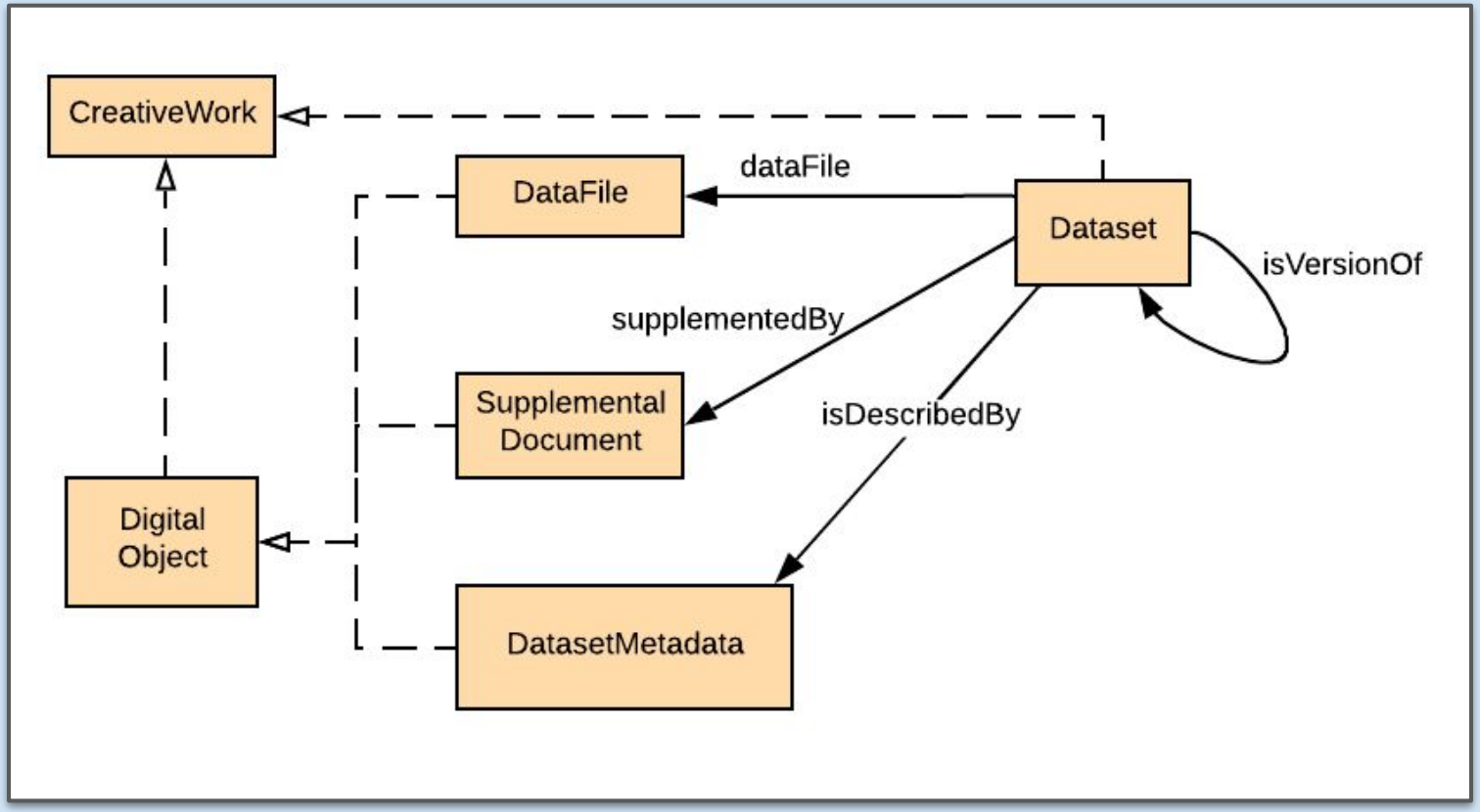


Can we operate on a Knowledge Graph?



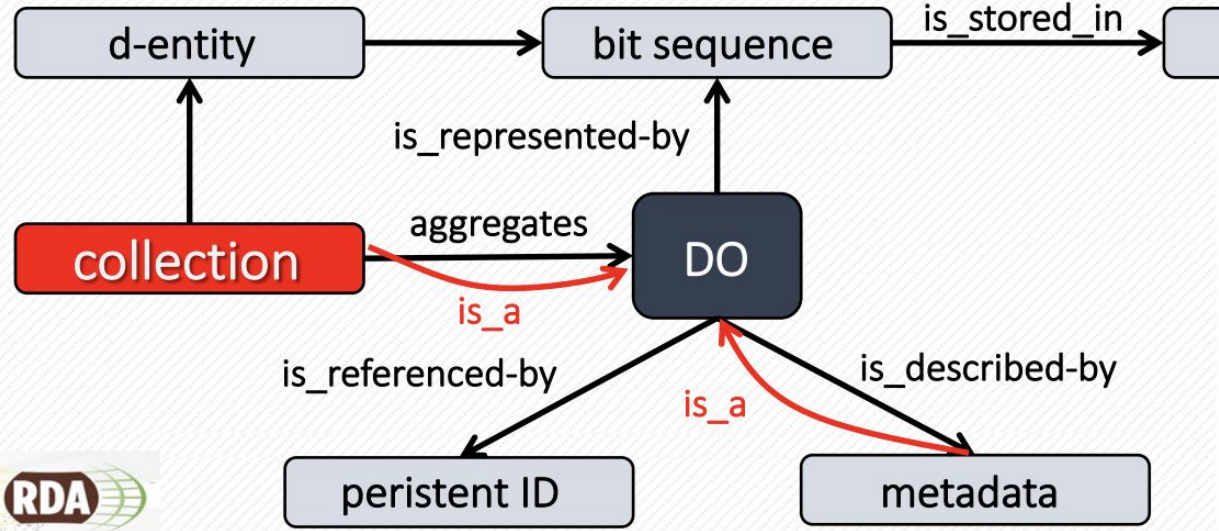
- Everything flow *through* the ontology
- Ocean Data Ontology
  http://schema.ocean-data.org
- Leverages our Data Expertise

Digital Object Model at BCO-DMO



Digital Object Model at DONA/CNRI

Digital Object pattern by CNRI/DONA

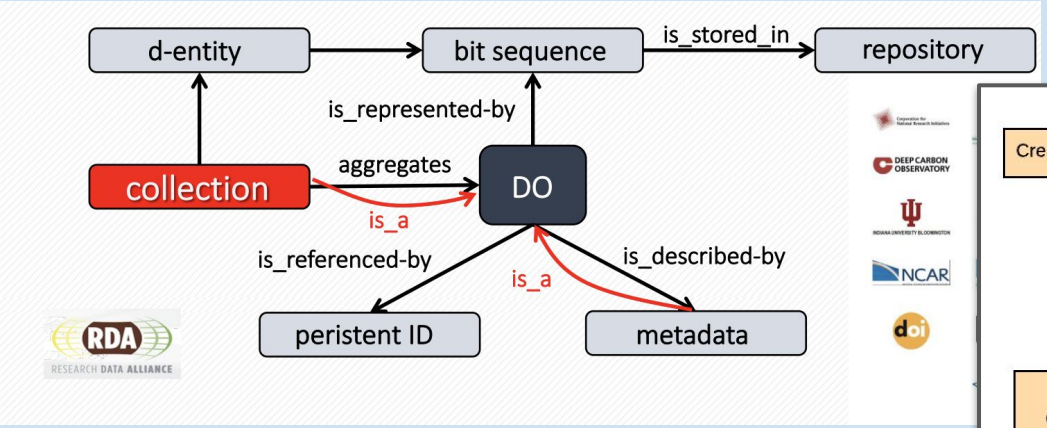


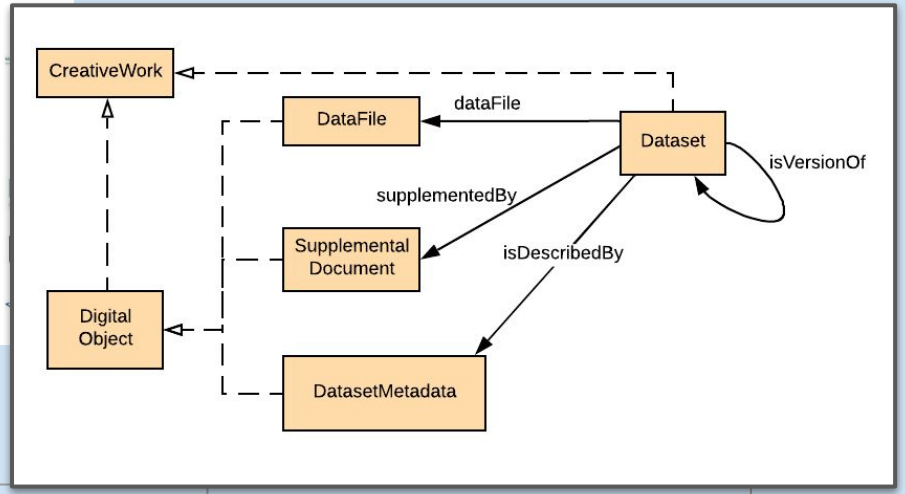
RESEARCH DATA ALLIANCE

DONA

Corporation for National Research Initiatives®

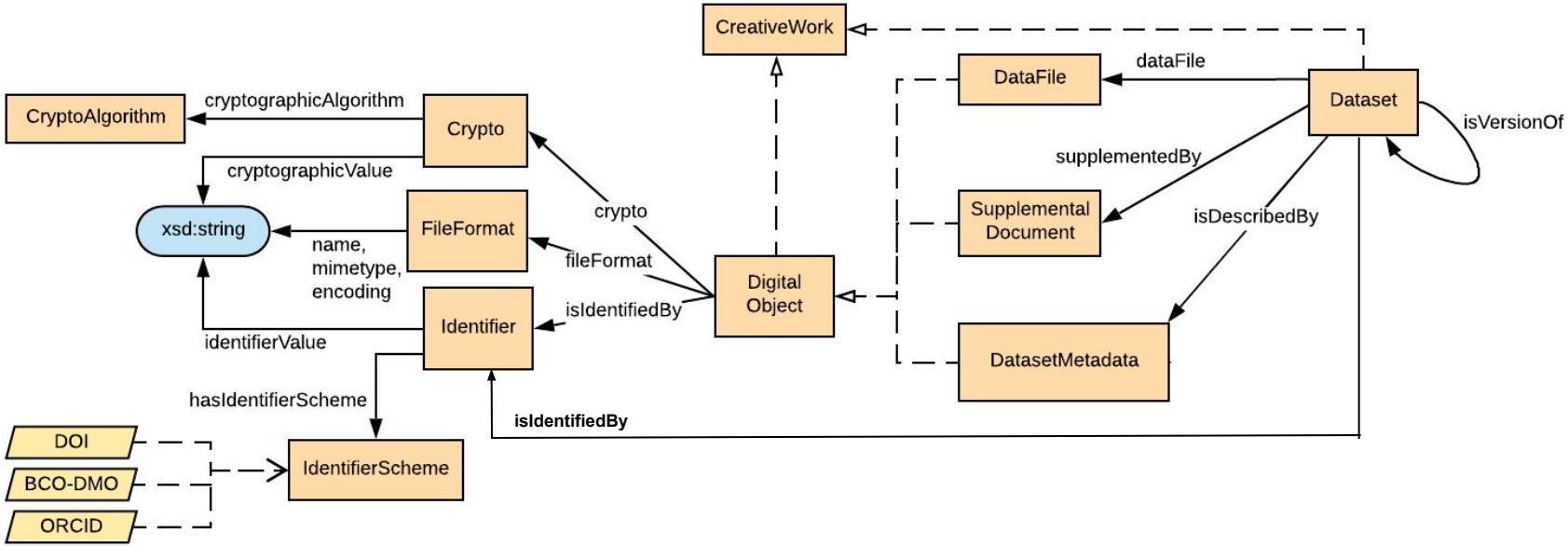
Aligning the Digital Object Models





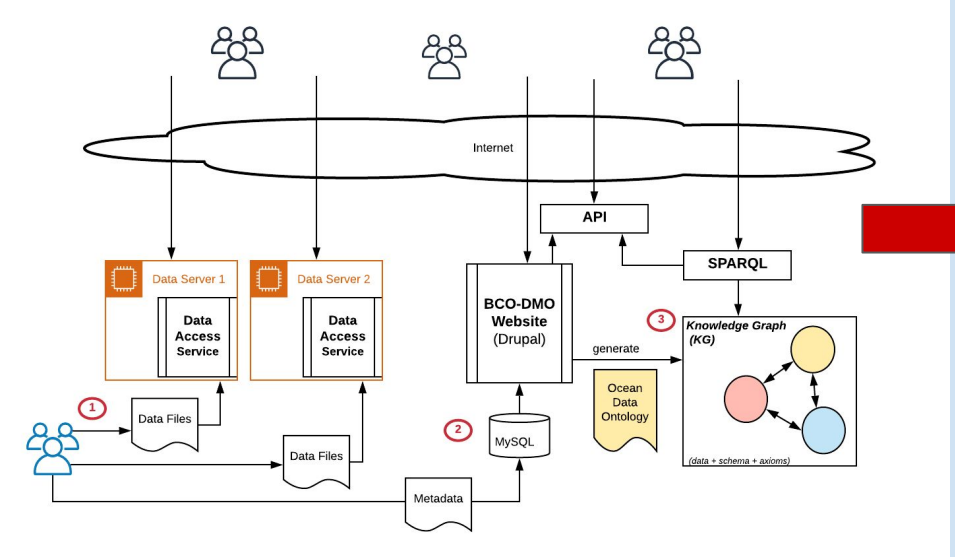
Ocean Data Ontology	***alignment***	**Digital Object Pattern**
Dataset	*is a*	Collection
dataFile supplementedBy isDescribedBy	*sub-property*	aggregates

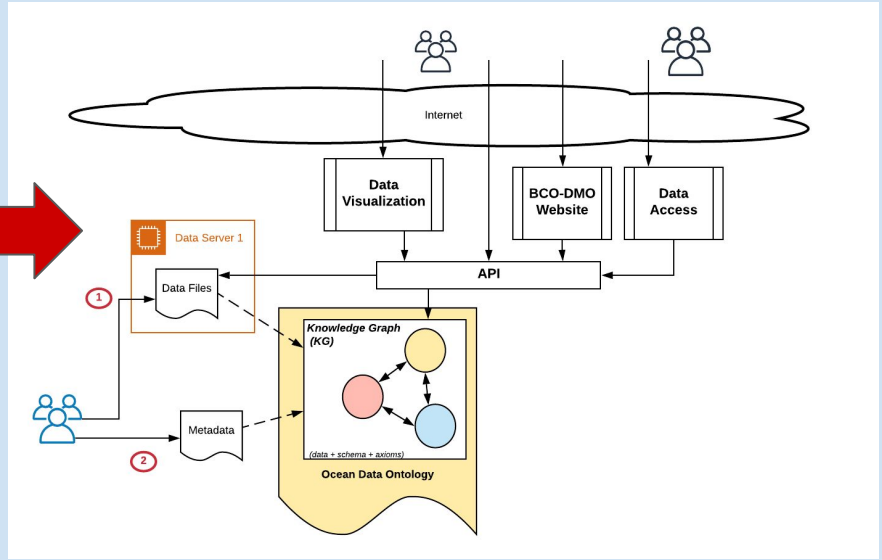
Implementing DONA Concepts



Use Case: Managing Access to Data

How do we address ***access*** to these digital objects?





Use Case: Managing Access to Data

PROBLEM: Access to Data was assumed to always be via one HTTP URL. Assumptions were made in code about this URL.

`**Dataset**` class had a property called `**hasDatasetURL**`
of primitive datatype `**xsd:anyURI**`.

odo:hasAward	<http://lod.bco-dmo.org/id/award/54626>
odo:hasBriefDescription	"Shipboard ADCP data from the US-GLOBEC Georges Bank Program study area."@en-us
odo:hasDatasetURL	"http://www.bco-dmo.org/dataset/2291/data"^^xsd:anyURI
odo:hasProcesingDescription	" The ADCP data are processed using the CO the University of Hawaii with additional capabilities developed at Brookhaven. "^^rdf:HTML
odo:osprey_page	<http://www.bco-dmo.org/dataset/2291>
odo:restricted	"0"^^xsd:boolean

Use Case: Managing Access to Data

Assumptions made about this Dataset URL:

#1 If the data were restricted/embargoed, a URL was not entered into this field.

Instead,

- ❏ the URL was pasted at the bottom of a administrative comments field
- ❏ when embargo was lifted, it was cut-n-pasted into place.

- ❏ There's a conflation between the ***absence of a URL*** *entirely* and the data being restricted.

Use Case: Managing Access to Data

Assumptions made about this Dataset URL:

#2 *If the URL contained a certain pattern*, it belonged to a specific data access application.

e.g. `/jg/serv/` in the URL was a proxy for other capabilities (subsetting, conversion, etc.)

In effect, no URL ever can accidentally contain this pattern

Use Case: Managing Access to Data

PROBLEMS of ASSUMPTIONS in SOFTWARE

***#1 Conflation* of the *meaning* of *absence* of *hasDatasetURL*.**

#2 Imbuing *meaning* on a *sequence of characters.*

GOAL: Develop access to digital objects that is extendible to any protocol ***afforded*** by some application.

Use Case: Managing Access to Data

What are the Competency Questions
that help us transform assumptions in code
into knowledgeable (smart) data?

Grüninger M., Fox M.S. (1995) The Role of Competency Questions in Enterprise Engineering. In: Rolstadås A. (eds) Benchmarking — Theory and Practice. IFIP Advances in Information and Communication Technology. Springer, Boston, MA, https://doi.org/10.1007/978-0-387-34847-6_3

Use Case: Managing Access to Data

Main Competency Questions

1. What actions can be taken on a Dataset's files from the Web?

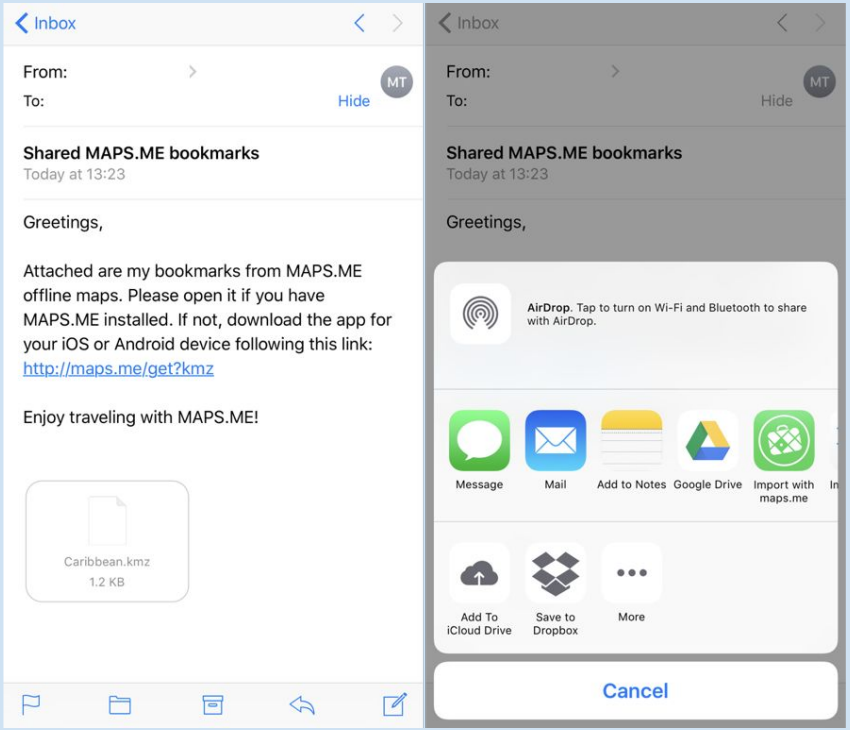
2. How would an application build a request?

3. What access policies govern these actions?
   aka Who is/isn't allowed to perform the action?

Use Case: Managing Access to Data

GOAL: Develop access to digital objects that is extendible to any protocol ***afforded*** by some application.

Inspired by the
Web Share API
&
Schema.org Actions

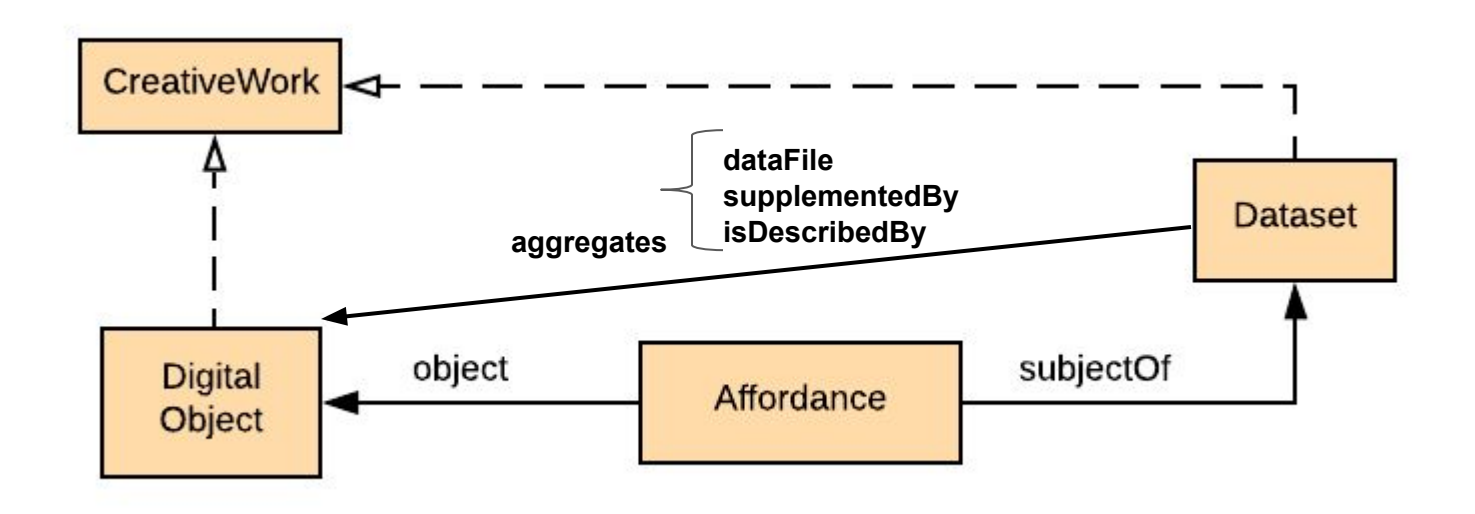


https://developers.google.com/web/updates/2016/09/navigator-share

https://developers.google.com/gmail/markup/actions/declaring-actions

Use Case: Managing Access to Data

An **Affordance** is some capability to access a DigitalObject scoped to a particular Dataset.

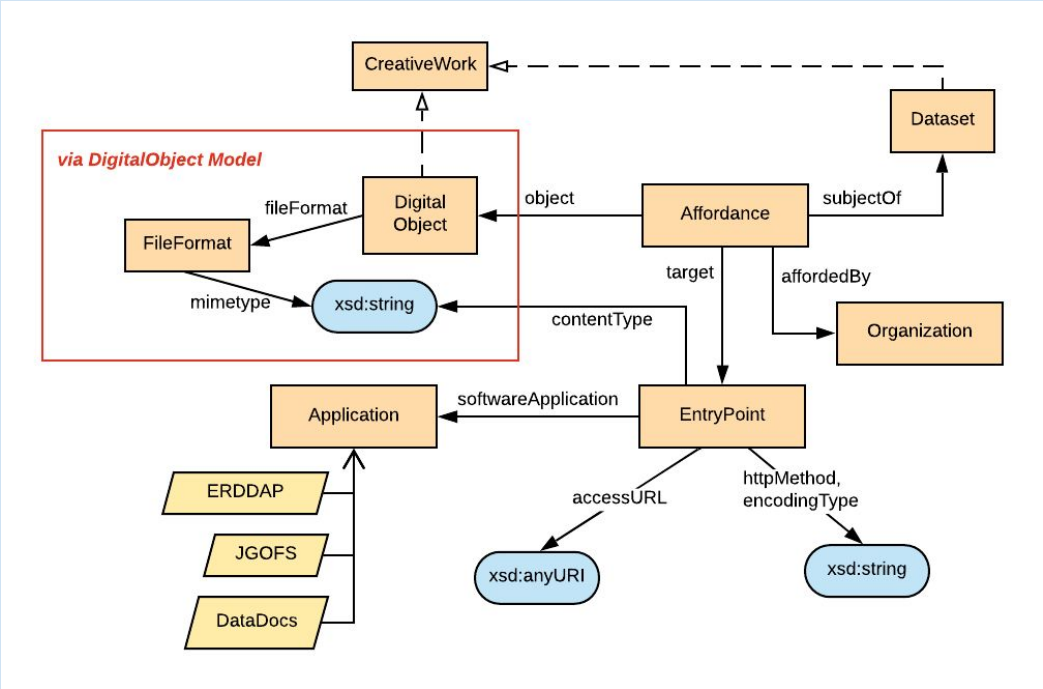


Why scoped to a Dataset?

DigitalObjects can be shared across Datasets. Scoping helps us control access (restricted data)

Use Case: Managing Access to Data

An **Affordance** directs you to an entry point is provided by some Organization some entry point from some software.



Use Case: Managing Access to Data

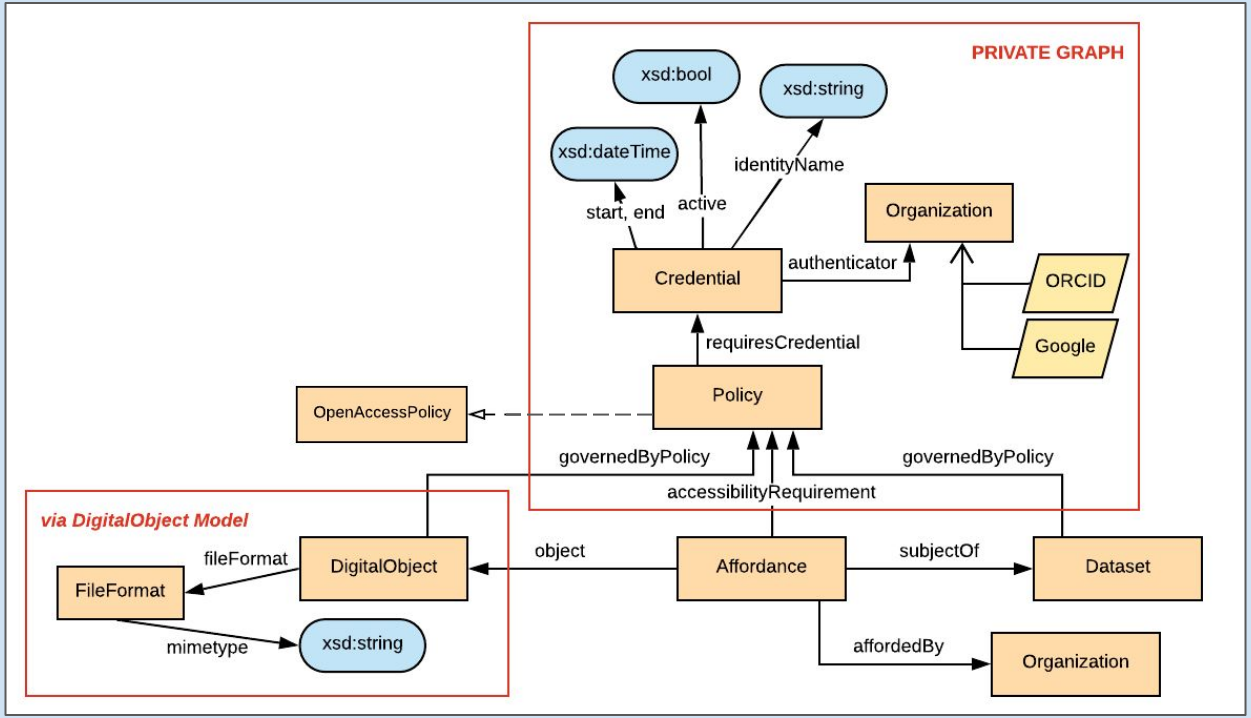
An **Affordance** may have a accessibility requirement…

1) explicitly set

2) defined by a Policy on the Digital Object
(for all its Affordances)

3) defined by a Policy on the Dataset
(for all Affordances of all its DOs)

4) Overriden by an `OpenAccessPolicy`



Inspired by the Web Server Access Control models

Use Case: Managing Access to Data

Main Competency Questions

1. What actions can be taken on a Dataset's files from the Web?

2. How would an application build a request?

3. What access policies govern these actions?
   aka Who is/isn't allowed to perform the action?

Use Case: Managing Access to Data

Main Competency Questions

1. What actions can be taken on a Dataset's files from the Web?

2. How would an application build a request?

3. What access policies govern these actions?
   aka Who is/isn't allowed to perform the action?

Competency Questions help you validate whether an ontology by turning each question into a SPARQL query. --

(Bezerra et al, 2013, DOI: 0.1109/WI-IAT.2013.199)

Use Case: Managing Access to Data

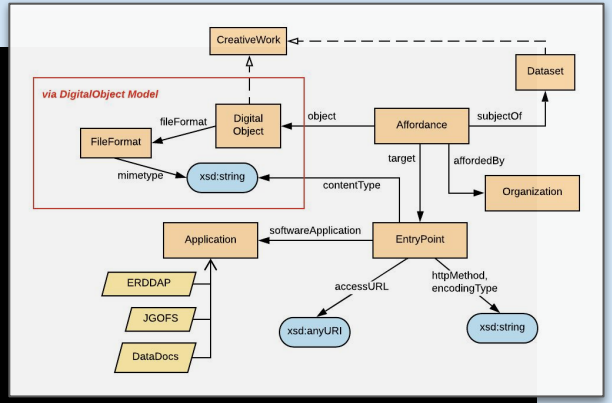
CQ1: **What actions can be taken on a Dataset's files from the Web?**

```
PREFIX odo: <http://ocean-data.org/schema/>

SELECT ?digitalobject ?app ?url

WHERE {
 ?affordance  odo:subjectOf   ?dataset .
 BIND(<http://lod.bco-dmo.org/id/dataset/2298> as ?dataset)

 ?affordance  odo:object      ?ditigalobject .
 ?affordance  odo:target      ?target .
 ?target      odo:softwareApplication   ?app .
 ?target      odo:accessURL   ?url
}
```



Use Case: Managing Access to Data

Competency Questions as SPARQL queries become the basis of your code.

```
function getAffordancesForDataset(string dataset_uri) {

  var query = """PREFIX odo: <http://ocean-data.org/schema/>
   SELECT ?digitalobject ?app
   WHERE {
    ?affordance  odo:subjectOf   ?dataset .
    BIND(:dataset_uri as ?dataset)

    ?affordance  odo:object      ?ditigalobject .
    ?affordance  odo:target      ?target .
    ?target      odo:softwareApplication   ?app .
    ?target      odo:accessURL   ?url
  }""";

  return execute(query);
}
```

Lessons Learned

1. It's easy for developers to encode knowledge in software.

2. Harder for software systems to share knowledge than it is for data. (Data is more resilient!)

3. Transform those assumptions into data,

4. using patterns that already work (if possible),

5. but, ALWAYS stay TRUE to your USE CASES.

Questions?

Thank you!

Adam Shepherd, *Technical Director*, BCO-DMO
Danie Kinkade, *Director*, BCO-DMO
Douglas Fils, *Data Manager*, Consortium of Ocean Leadership

BCO-DMO: https://www.bco-dmo.org
Ocean Data Ontology: http://ocean-data.org/schema/

BCO-DMO

NSF

Woods Hole Oceanographic Institution

Extra Slides

Use Case: Provenance of Data Curation

What steps were taken to alter the original data submission?

- ❑ **Reproducibility** *(Research Community)*
- ❑ **Transparency** *(General Public and our Funders)*
- ❑ **Error Management** *(For Ourselves)*

Use Case: Provenance of Data Curation

best_hit_annotation	best_hit_taxon_id	st1_050m	st1_090m	st1_120m	st1_200m	st1_300m	st1_400m	st1_600m	st3_040m	st3_060m	st3_120m	st3_1
nitrate reductase alpha subunit	247490	0	0	0	0	122	121	116	0	0	17	
nxrB1; putative nitrate oxidoreductase subunit beta (E	330214	0	0	0	1	136	173	153	0	0	18	
groEL; chaperonin GroEL (EC:3.6.4.9); K04077 chapero	167546	80	91	59	35	2	2	1	60	44	24	
ccmK; carboxysome shell protein CsoS1	167555	155	162	94	38	0	0	0	54	40	39	
putative UreA ABC transporter; substrate binding prot	167546	202	203	169	158	26	30	19	100	86	27	
ligand-binding protein; OpuAC family	859653	7	7	8	7	51	69	74	3	1	19	
ABC transporter	314261	17	20	19	9	30	35	36	12	15	22	
ABC transporter; substrate-binding protein; family 5	89187	0	0	0	0	50	50	65	0	0	2	
glnA; glutamine synthetase; glutamate--ammonia ligas	146891	62	60	54	58	3	4	2	60	53	4	
amino acid ABC transporter substrate-binding protein	913324	3	2	2	4	33	78	43	0	0	6	
F0F1 ATP synthase subunit beta	93058	57	63	76	34	5	4	3	39	40	47	
peptide ABC transporter; periplasmic substrate-bindin	375451	0	2	3	0	41	48	44	0	0	6	
glutamate/glutamine/aspartate/asparagine ABC trans	488538	2	7	11	2	31	52	44	2	3	5	
hypothetical protein	1090946	1	1	7	0	51	47	37	0	0	5	
ABC transporter binding protein	859653	88	68	33	29	1	4	3	38	32	10	
rbcL; ribulose bisophosphate carboxylase; K01601 ribu	146891	46	57	40	36	1	3	0	37	41	15	
nd	1073573	20	10	2	1	26	44	29	10	8	2	

What does `st1_050m` mean?

An observation made at the location called station '1'
at a depth of 50 meters

Use Case: Provenance of Data Curation

BCO-DMO Data Managers split these observations apart.

Data Curation occurs using Declarative Workflows

	station	depth	spectral_count
6	8	200	0
6	9	40	12
6	9	70	4
6	9	380	0
1	12	40	0
1	12	120	0
1	12	300	0
4	1	50	0
4	1	90	0
4	1	120	0
4	1	200	0
4	1	300	0
4	1	400	0
1	1	600	0

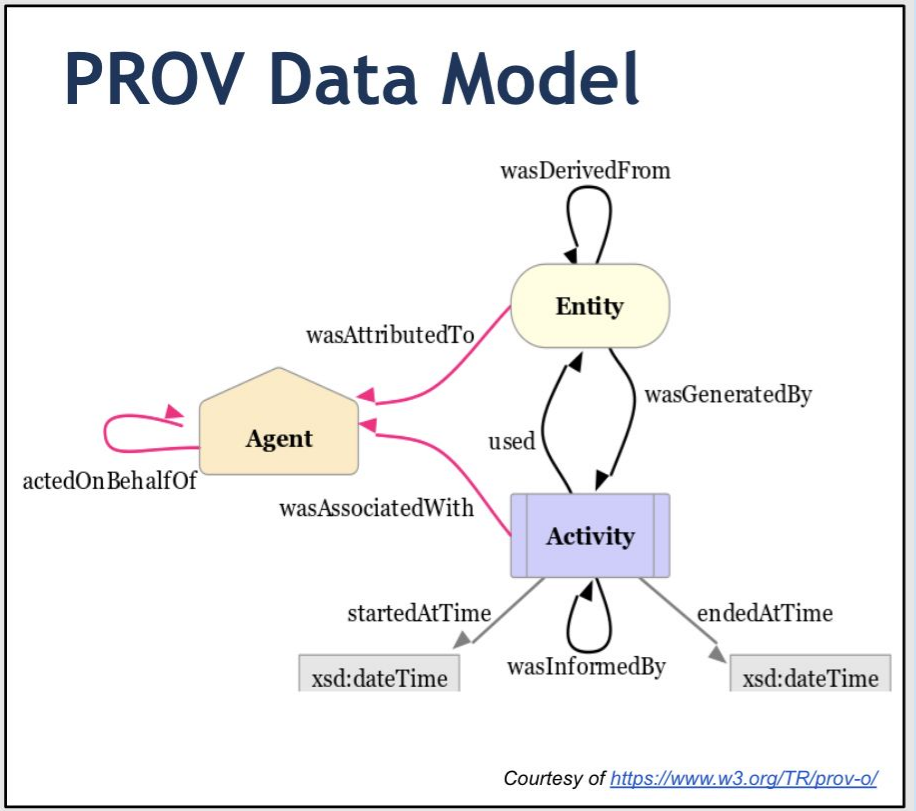
Use Case: Provenance of Data Curation

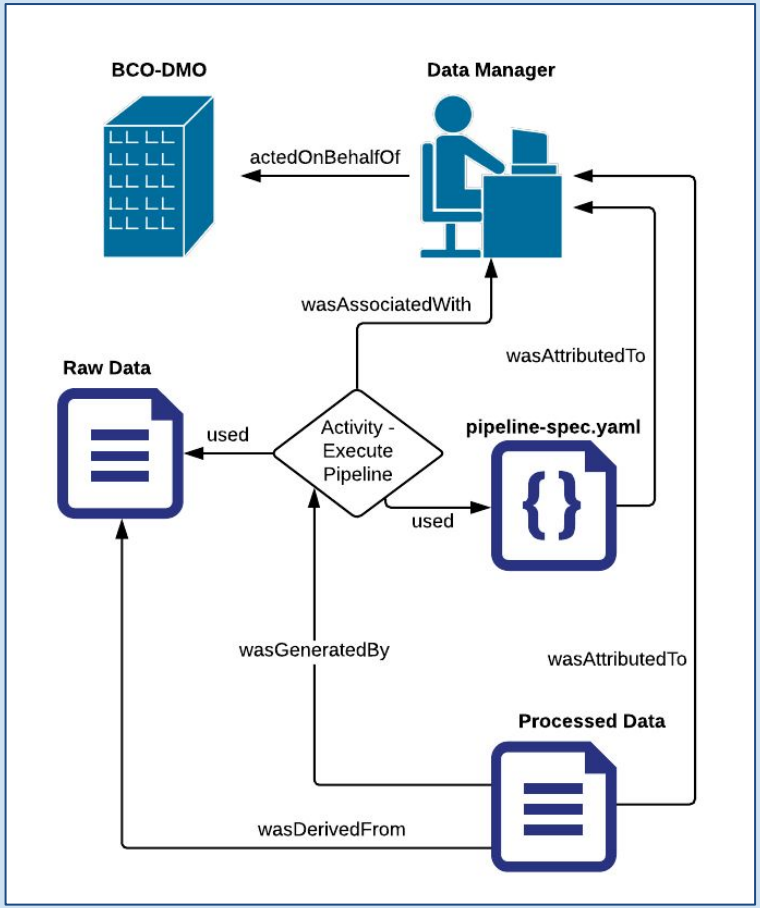
pipeline-spec.yaml

```
- run: bcodmo_pipeline_processors.convert_to_decimal_degrees
  cache: True
  parameters:
    resources: [mcmurdo_epifauna]
    fields:
    - {input_field: lat_long, format: degrees-decimal_minutes, output_field: lat_converted, directional: '',
      pattern: "(?P<degrees>.*)\xB0 (?P<decimal_minutes>.*)'(?P<directional>.)\\ .*\xB0 .*'."}
- run: bcodmo_pipeline_processors.convert_to_decimal_degrees
  cache: true
  parameters:
    resources: [mcmurdo_epifauna]
    fields:
    - {input_field: lat_long, format: degrees-decimal_minutes, output_field: long_converted, directional: '',
      pattern: ".*\xB0 .*'. (?P<degrees>.*)\xB0 (?P<decimal_minutes>.*)'(?P<directional>.)"}
```

frictionlessdata.io
github.com/frictionlessdata/datapackage-pipelines

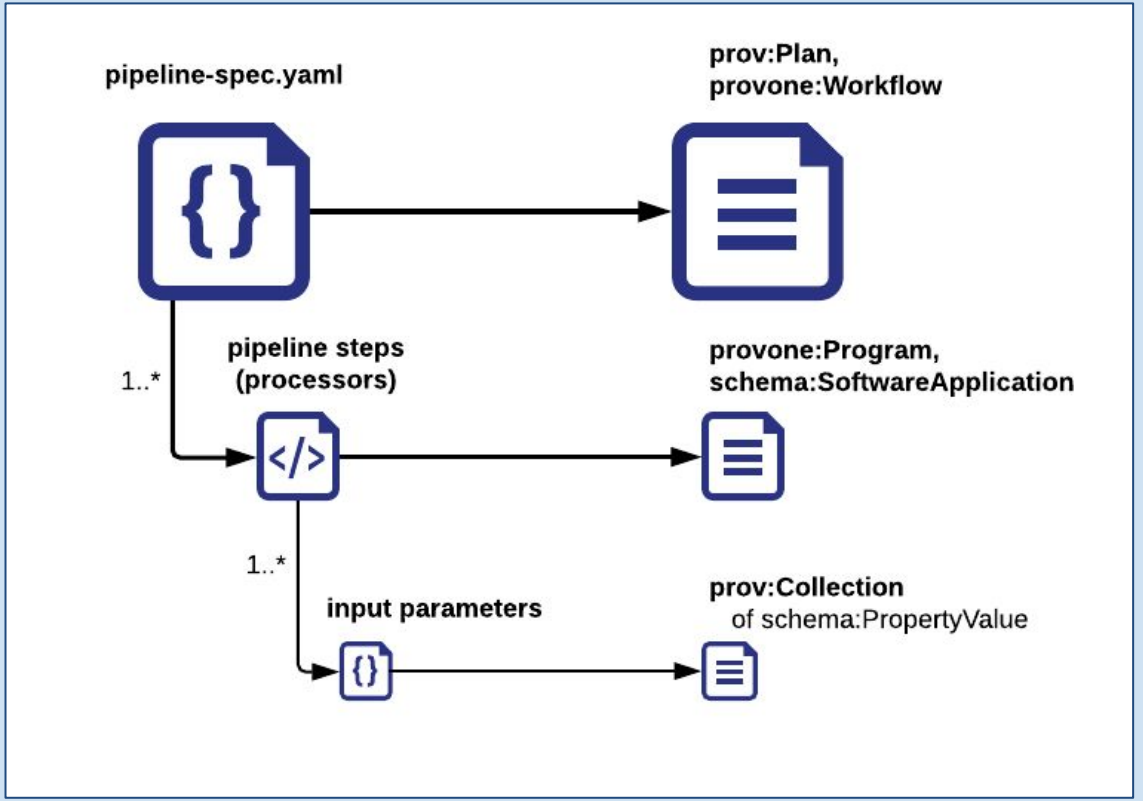
Use Case: Provenance of Data Curation





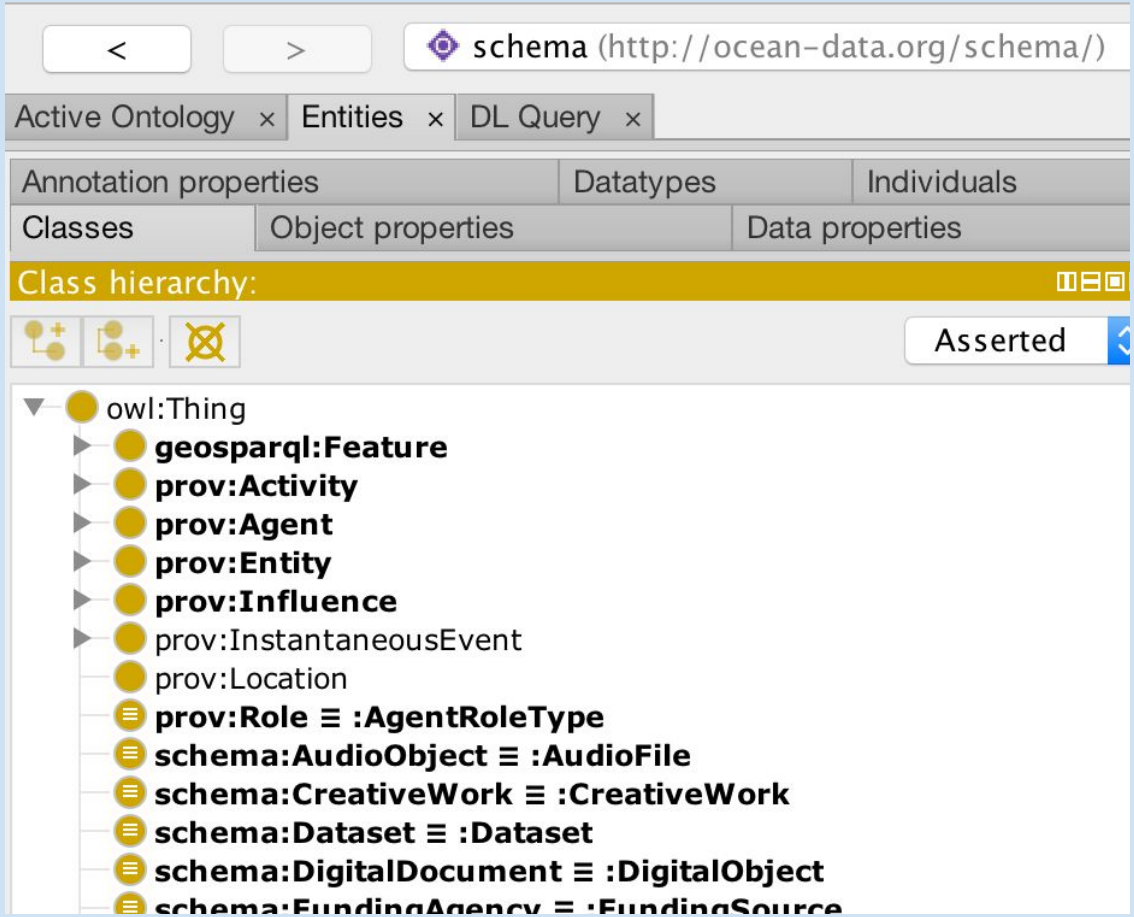
Use Case: Provenance of Data Curation

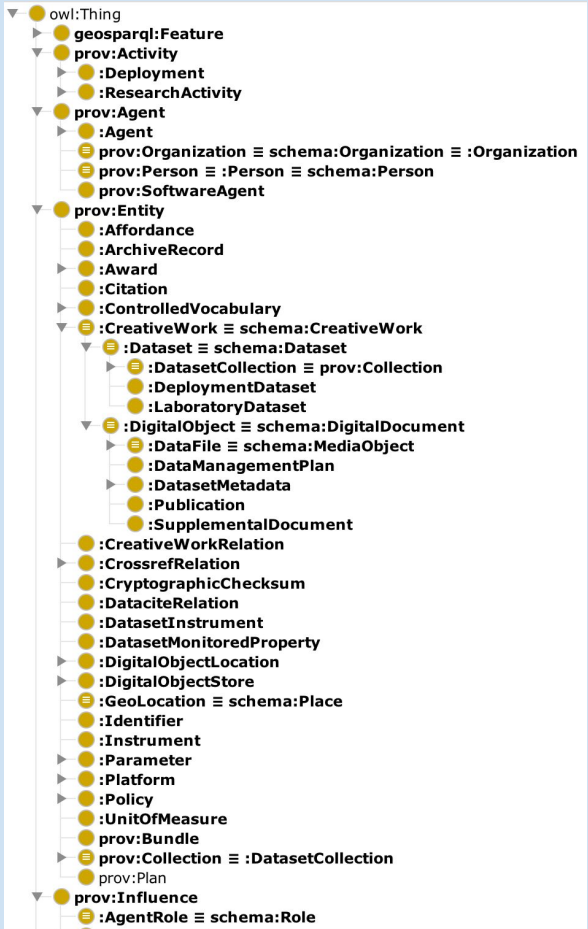
Mapping elements of
a data file transform
pipeline to
PROV Data Model
w. some Schema.org



prov: http://www.w3.org/ns/prov#
provone: http://purl.dataone.org/provone/2015/01/15/ontology#
schema: http://schema.org/

Use Case: Provenance of Data Curation





Use Case: Provenance of Data Curation