

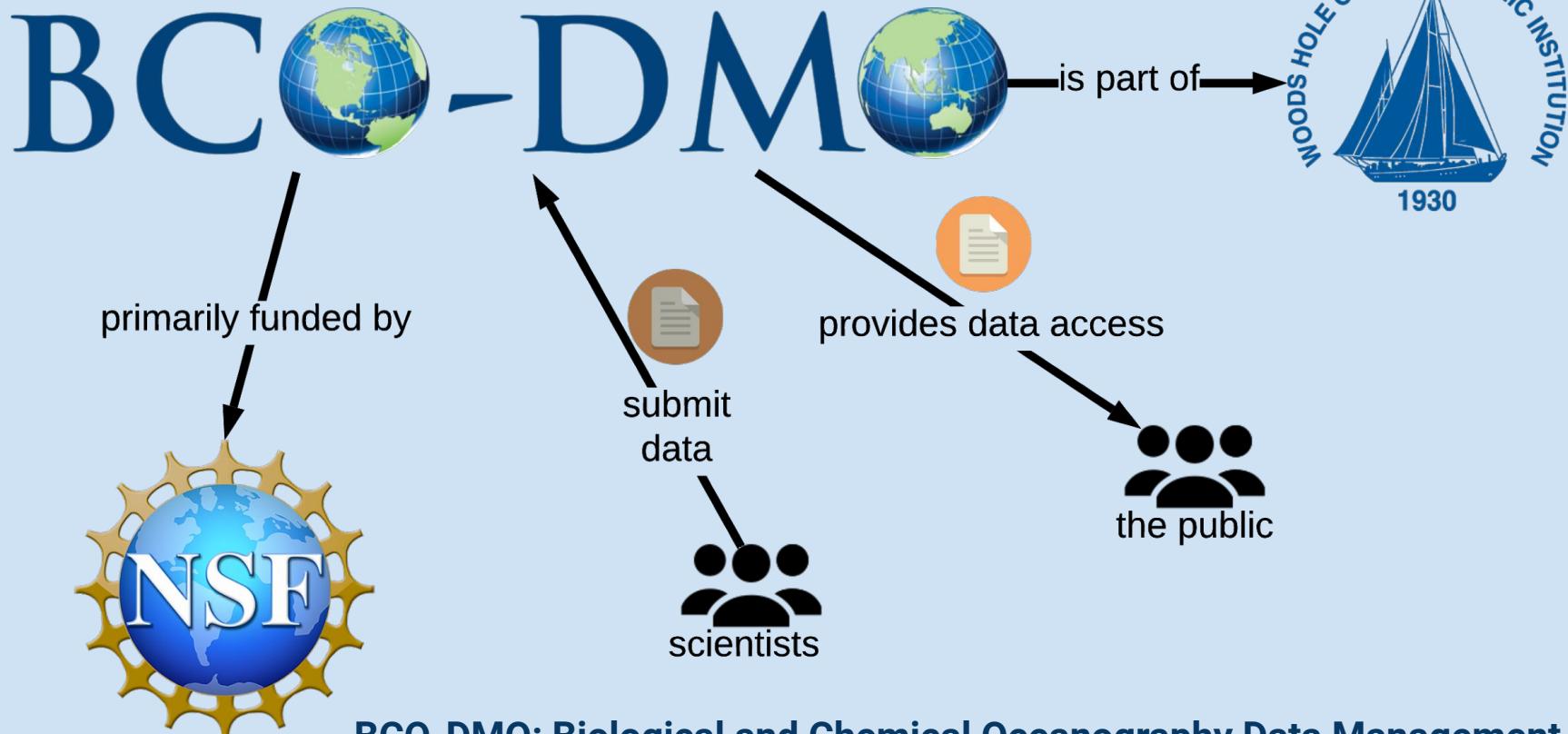
Reconstituting a Digital Repository through Use Case Driven Ontology Development

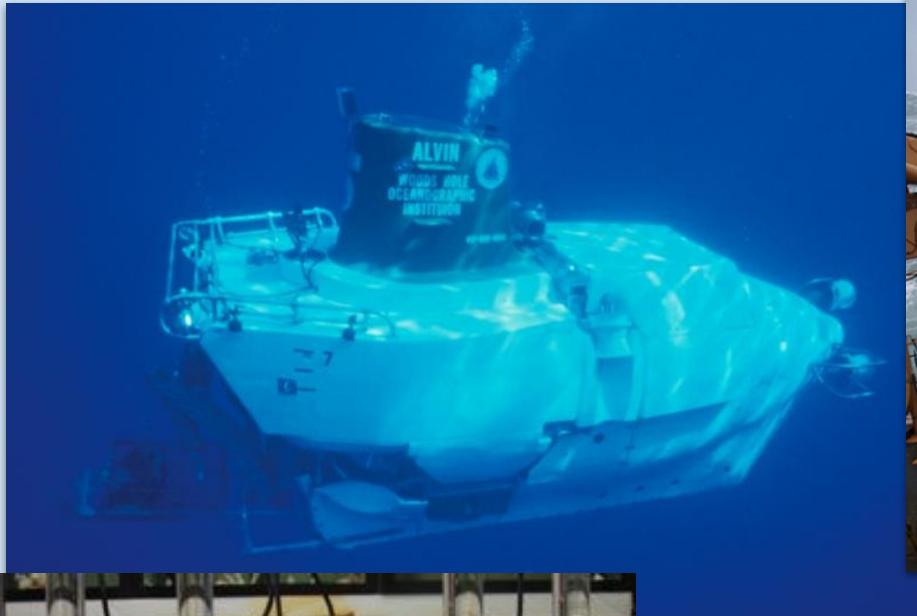
SIG-CM Workshop
JCDL 2019 June 6th, 2019

Adam Shepherd, Danie Kinkade, Douglas Fils



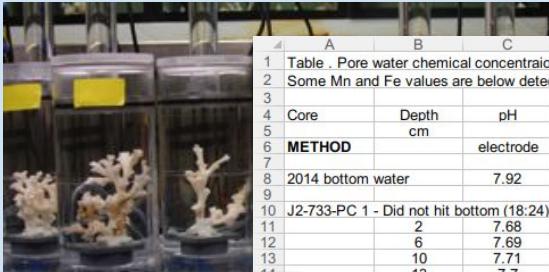
Who are we?







Did you record the
metadata?



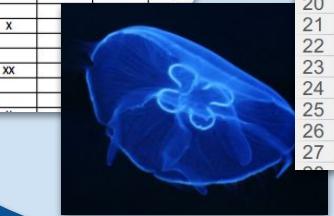
A	B	C	D
1	Table . Pore water chemical concentration data and lo		
2	Some Mn and Fe values are below detection and are		
3			
4	Core	Depth	pH
5		cm	alkalinity
6	METHOD		mmol/kg
7		electrode	titration
8	2014 bottom water	7.92	2.32
9			
10	J2-733-PC 1 - Did not hit bottom (18:24) and was pos		
11	2	7.68	2.08
12	6	7.69	2.11
13	10	7.71	2.20
14	13	7.7	2.22
15	16	7.69	2.22
16	18	7.71	2.22
17			
18	-733-PC 2 - Did not hit bottom (18:27) and was positioned next to PC 1. 22.0		
19	2	7.69	2.13
20	7	7.70	21.9
21	11	7.73	548.1
22	15	7.70	
23	18	7.72	
24	20	7.70	
25	22	7.72	
26			
27	-733-PC 4 - Hit bottom (18:01) and wa		
28	3	7.67	
29	6	7.69	
30	8	7.71	
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			
50			
51			
52			
53			
54			
55			
56			
57			
58			
59			
60			
61			
62			
63			
64			
65			
66			
67			
68			
69			
70			
71			
72			
73			
74			
75			
76			
77			
78			
79			
80			
81			
82			
83			
84			
85			
86			
87			
88			
89			
90			
91			
92			
93			
94			
95			
96			
97			
98			
99			
100			
101			
102			
103			
104			
105			
106			
107			
108			
109			
110			
111			
112			
113			
114			
115			
116			
117			
118			
119			
120			
121			
122			
123			
124			
125			
126			
127			
128			
129			
130			
131			
132			
133			
134			
135			
136			
137			
138			
139			
140			
141			
142			
143			
144			
145			
146			
147			
148			
149			
150			
151			
152			
153			
154			
155			
156			
157			
158			
159			
160			
161			
162			
163			
164			
165			
166			
167			
168			
169			
170			
171			
172			
173			
174			
175			
176			
177			
178			
179			
180			
181			
182			
183			
184			
185			
186			
187			
188			
189			
190			
191			
192			
193			
194			
195			
196			
197			
198			
199			
200			
201			
202			
203			
204			
205			
206			
207			
208			
209			
210			
211			
212			
213			
214			
215			
216			
217			
218			
219			
220			
221			
222			
223			
224			
225			
226			
227			
228			
229			
230			
231			
232			
233			
234			
235			
236			
237			
238			
239			
240			
241			
242			
243			
244			
245			
246			
247			
248			
249			
250			
251			
252			
253			
254			
255			
256			
257			
258			
259			
260			
261			
262			
263			
264			
265			
266			
267			
268			
269			
270			
271			
272			
273			
274			
275			
276			
277			
278			
279			
280			
281			
282			
283			
284			
285			
286			
287			
288			
289			
290			
291			
292			
293			
294			
295			
296			
297			
298			
299			
300			
301			
302			
303			
304			
305			
306			
307			
308			
309			
310			
311			
312			
313			
314			
315			
316			
317			
318			
319			
320			
321			
322			
323			
324			
325			
326			
327			
328			
329			
330			
331			
332			
333			
334			
335			
336			
337			
338			
339			
340			
341			
342			
343			
344			
345			
346			
347			
348			
349			
350			
351			
352			
353			
354			
355			
356			
357			
358			
359			
360			
361			
362			
363			
364			
365			
366			
367			
368			
369			
370			
371			
372			
373			
374			
375			
376			
377			
378			
379			
380			
381			
382			
383			
384			
385			
386			
387			
388			
389			
390			
391			
392			
393			
394			
395			
396			
397			
398			
399			
400			
401			
402			
403			
404			
405			
406			
407			
408			
409			
410			
411			
412			
413			
414			
415			
416			
417			
418			
419			
420			
421			
422			
423			
424			
425			
426			
427			
428			
429			
430			
431			
432			
433			
434			
435			
436			
437			
438			
439			
440			
441			
442			
443			
444			
445			
446			
447			
448			
449			
450			
451			
452			
453			
454			
455			
456			
457			
458			
459			
460			
461			
462			
463			
464			
465			
466			
467			
468			
469			
470			
471			
472			
473			
474			
475			
476			
477			
478			
479			
480			
481			
482			
483			
484			
485			
486			
487			
488			
489			
490			
491			
492			
493			
494			
495			

A	B	C	D
1	Table . Pore water chemical concentration data and lo		
2	Some Mn and Fe values are below detection and are		
3			
4	Core	Depth	pH
5	cm		alkalinity
6	METHOD	electrode	mmol/kg 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
14	13	7.7	2.22
15	16	7.69	2.22
16	18	7.71	2.22



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

C	D	E	F	G	H
Video ID	Time (minute)	E	D	C.T.	C.O.
Clip015	4.06				7.73
	13.04	X		X	
	26.23			X	
	30.17		X		
	45.09	X			
	134.06			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

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

Dataset landing page

BCO-DMO Biological & Chemical Oceanography Data Management Office

DATA **RESOURCES** **ABOUT US** **Enter search terms**

DATABASE

- Programs 43
- Projects 1,045
- Deployments 2,835
- Platforms 594
- Datasets 9,410
- Instruments 480
- Parameters 1,415
- People 2,664
- Affiliations 583
- Funding 93
- Awards 1,966

Dataset: Water Chemistry

Cite This Dataset

Spatial Extent: N 22.8070 E -46.11082 S 22.8204 W -46.11083

Temporal Extent: 2014-04-11

Project: | |

Principal Investigator: These kids are jumping in the trash! These kids are jumping in! These kids are jumping in the trash!

BCO-DMO Data Manager: Shannon Rauch (Woods Hole Oceanographic Institution, WHOI BCO-DMO)

Version Date: 2019-04-11

Restricted: No

Validated: Yes

Current Status: Finalized and data-reviewed

GEOSPATIAL ACCESS

Why we needed a new infrastructure?

❑ Big Data challenges

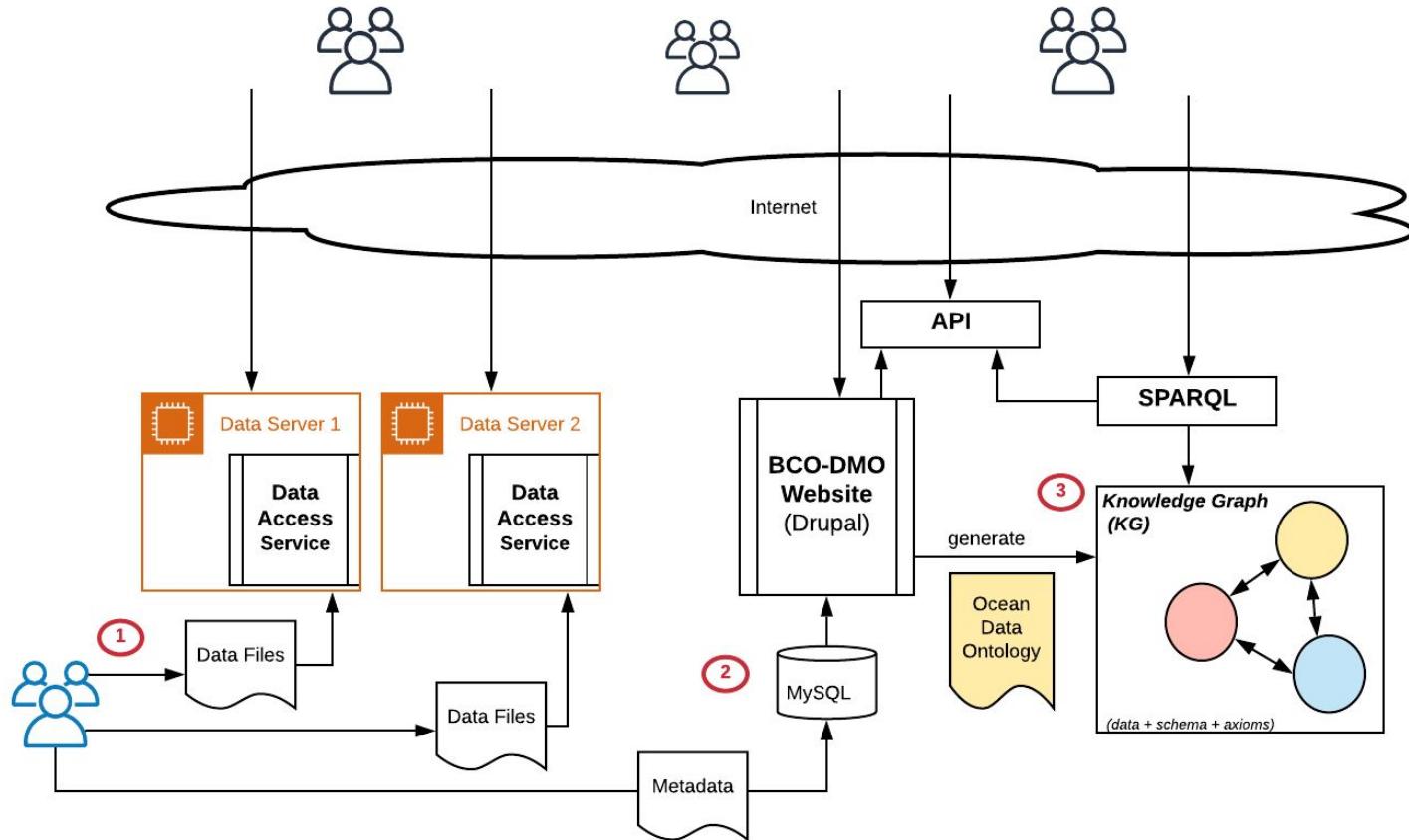
- ❑ Some data may be stored in other other external repositories
- ❑ But, our infrastructure depends on local management

❑ Software ages faster than data

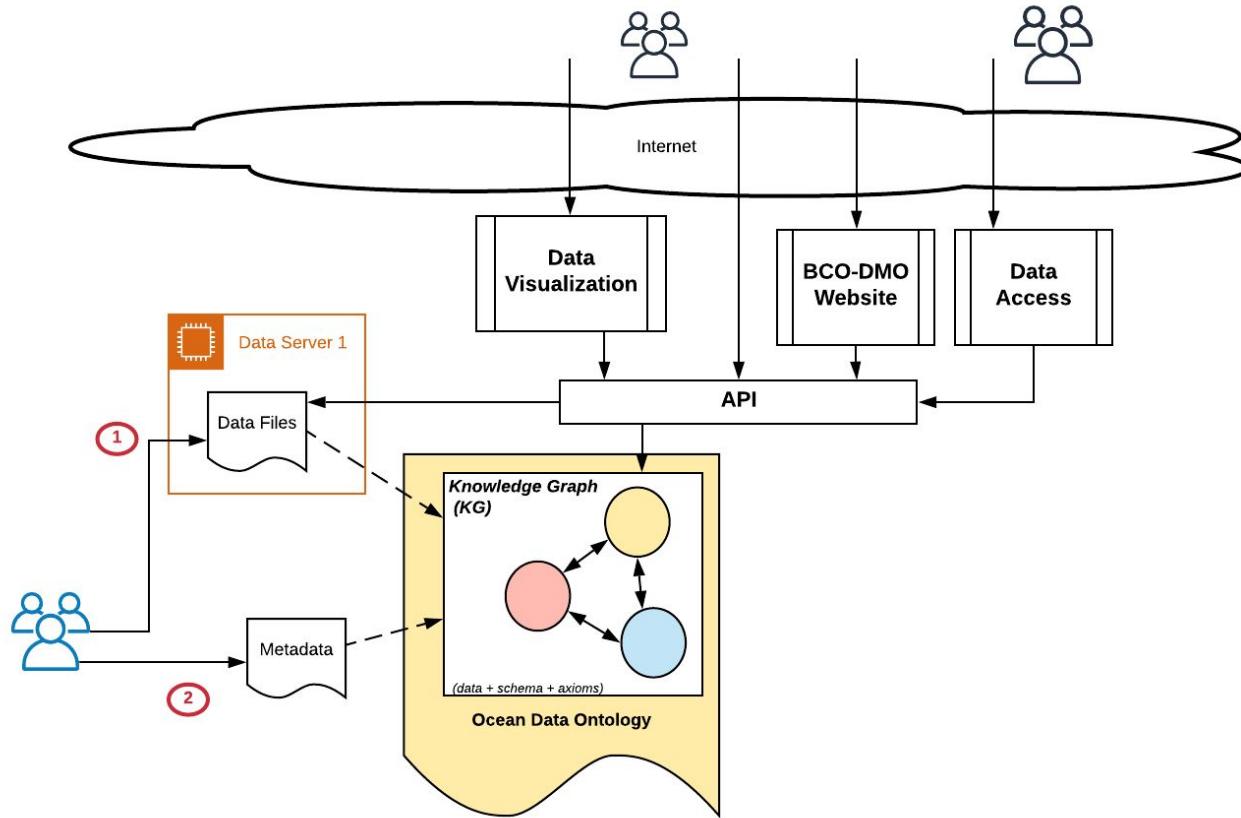
Data Submission
Metadata Catalog
Data Access
Data Visualization

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

OLD: Flat Architecture = Dependency

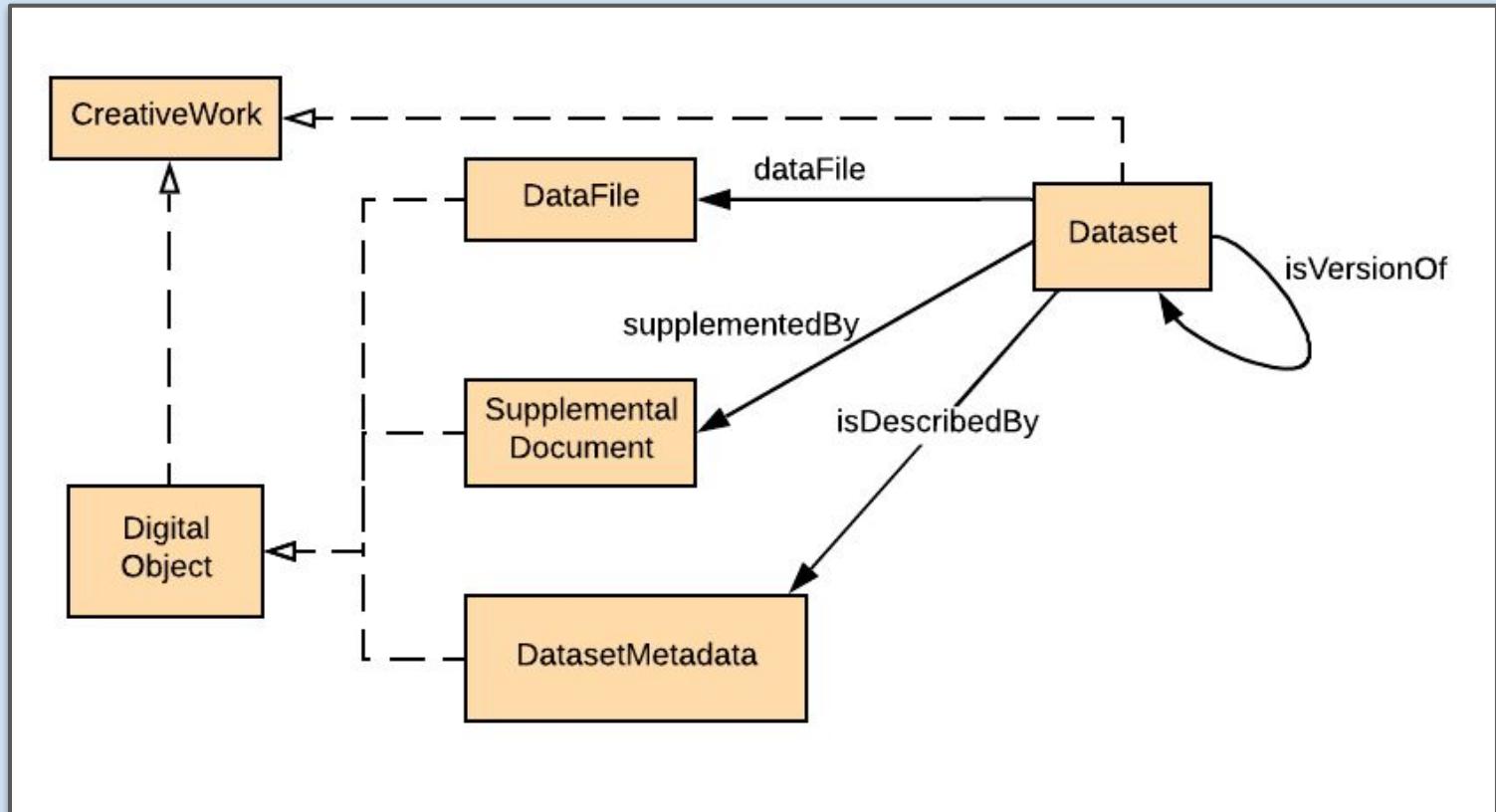


Can we operate on a Knowledge Graph?



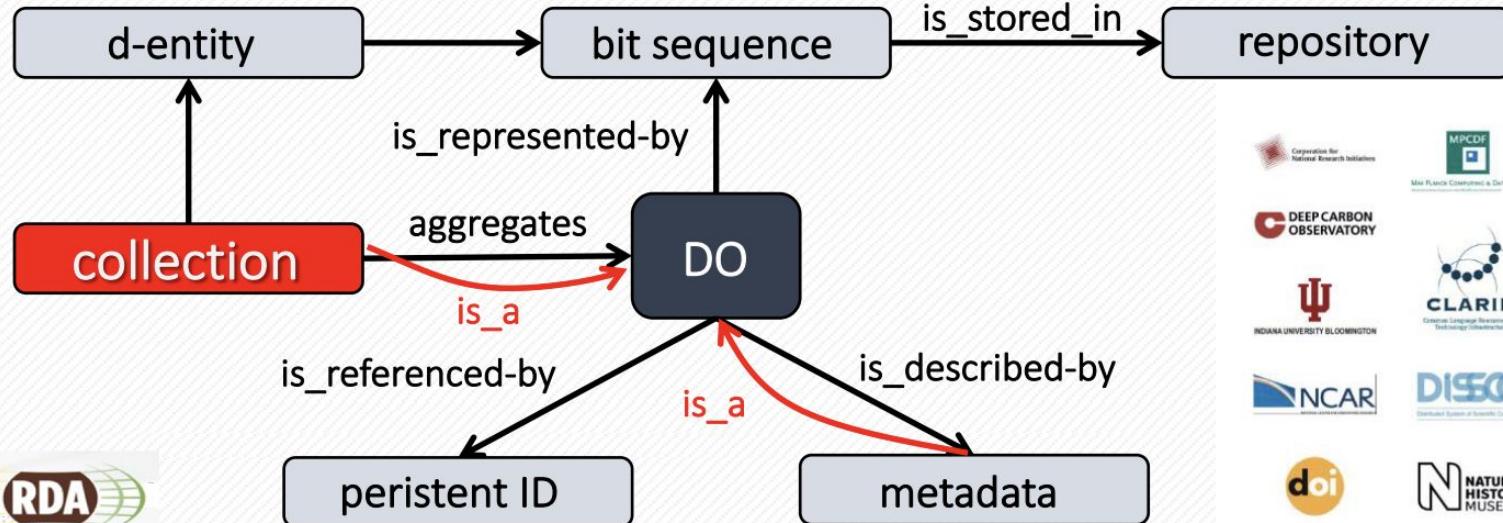
- Everything flow through the ontology
- Ocean Data Ontology
<http://schema.ocean-data.org>
- Leverages our Data Expertise

Digital Object Model at BCO-DMO

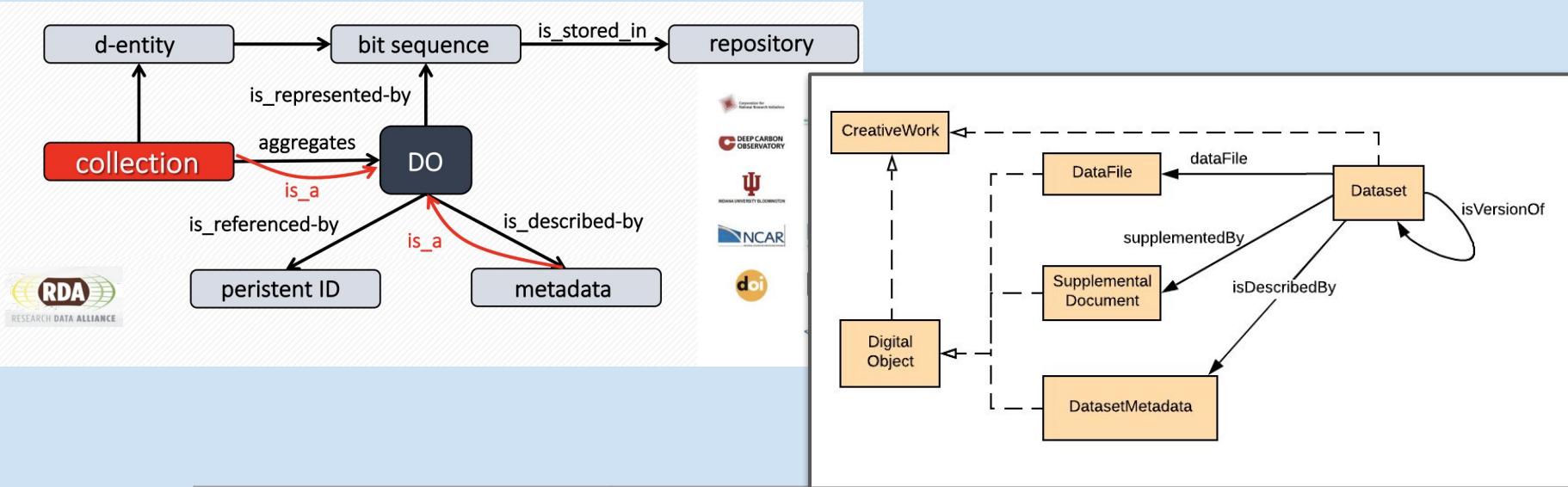


Digital Object Model at DONA/CNRI

Digital Object pattern by CNRI/DONA

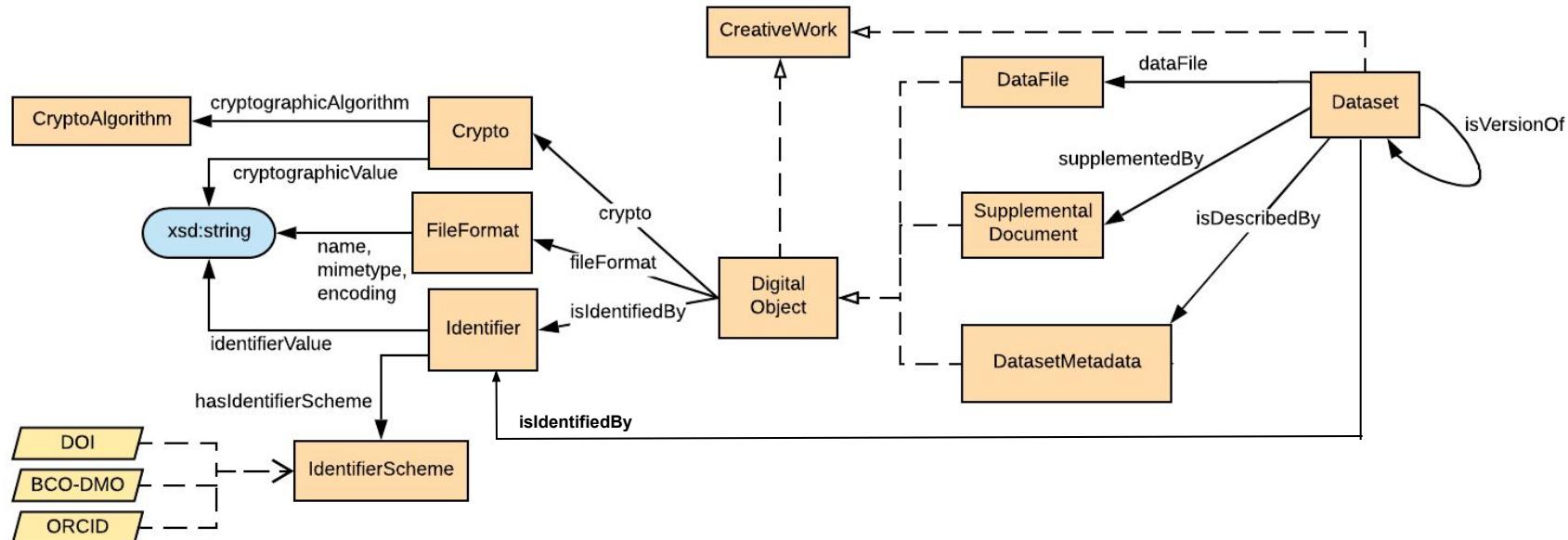


Aligning the Digital Object Models



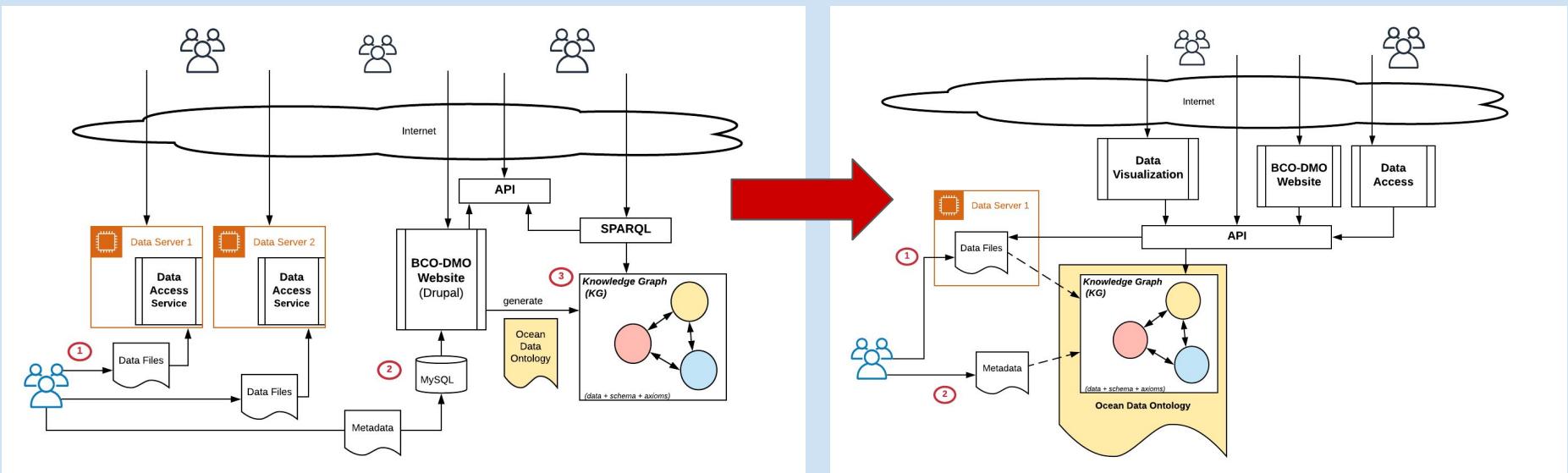
Ocean Data Ontology	<i>alignment</i>	Digital Object Pattern
Dataset	<i>is a</i>	Collection
dataFile supplementedBy isDescribedBy	<i>sub-property</i>	aggregates

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:hasProcessingDescription	"<div xmlns="http://www.w3.org/1999/xhtml" lang="en"><p>The ADCP data are processed using the CC the University of Hawaii with additional capabilities developed at Brookhaven.</p></div>"^^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

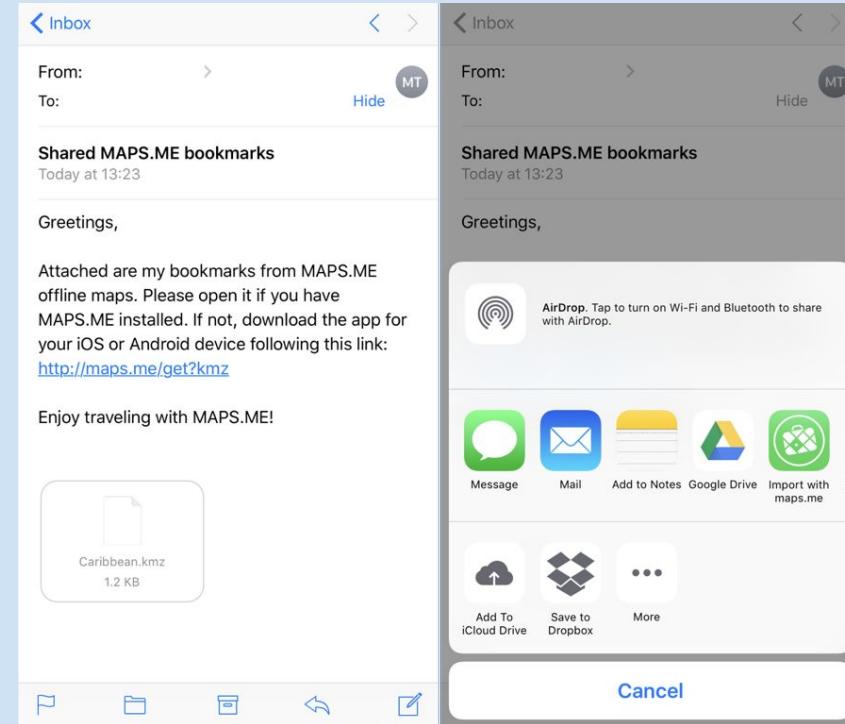
Main Competency Questions

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

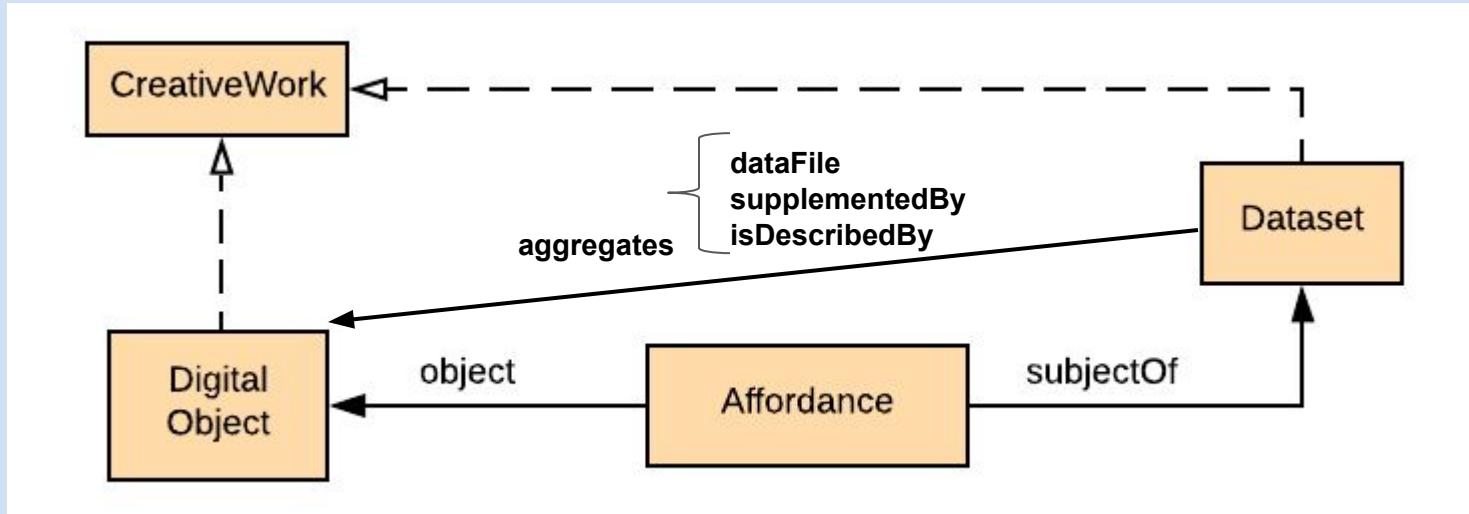


<https://developers.google.com/web/updates/2016/09/navigator-share>

<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.

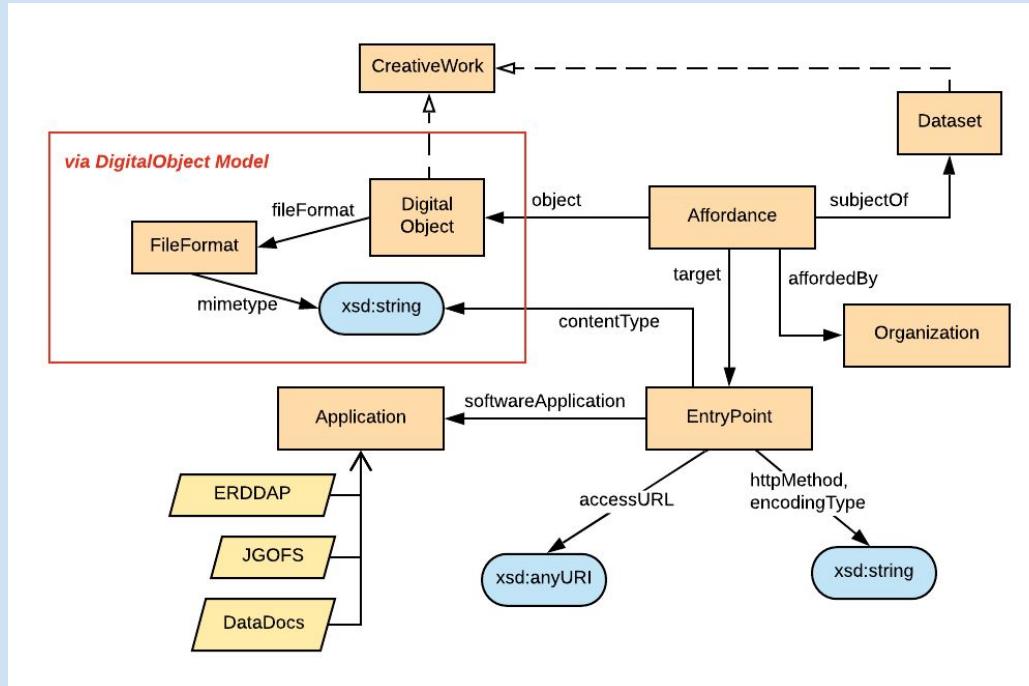


Why scoped to a Dataset?

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

Use Case: Managing Access to Data

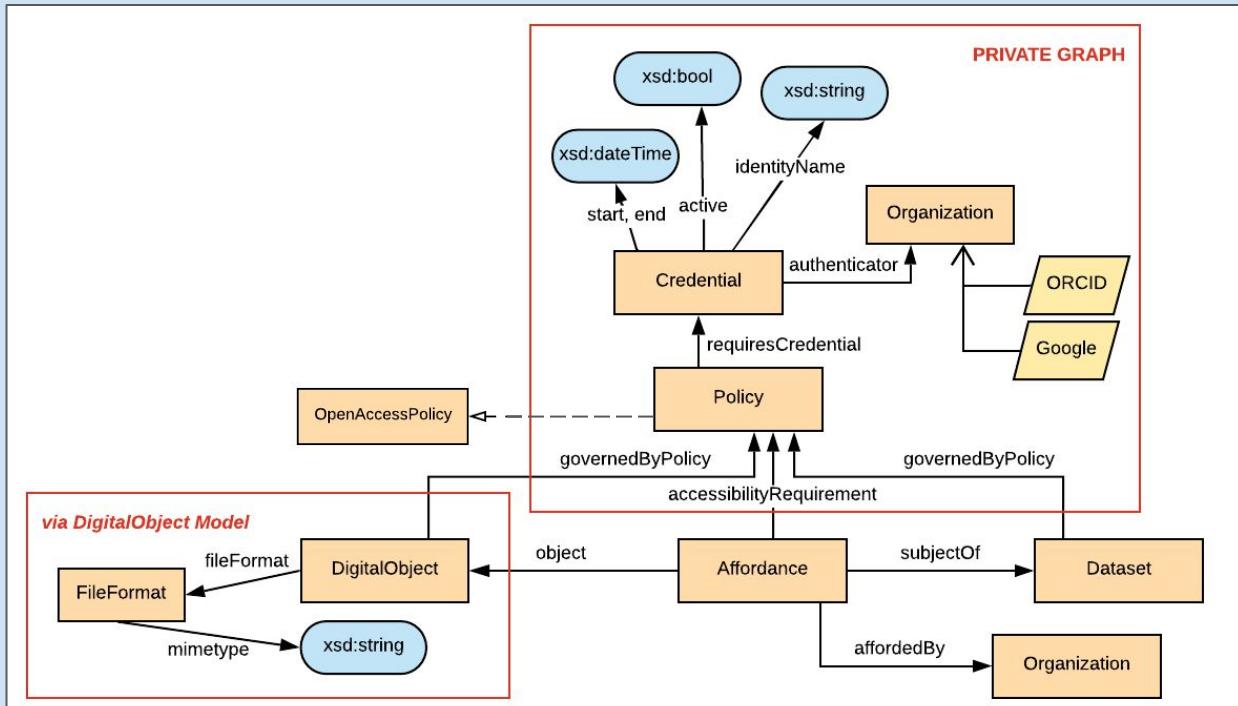
An **Affordance** directs you to an entry point provided by some Organization some entry point from some software.



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)
```

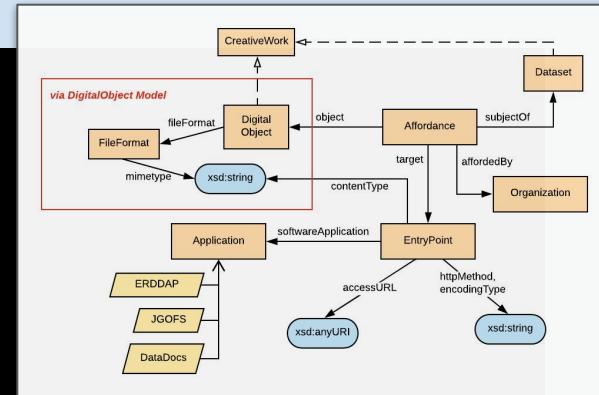
```
    ?affordance odo:object ?ditigalobject .
```

```
    ?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/>



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_180m
nitrate reductase alpha subunit	247490	0	0	0	0	122	121	116	0	0	17	1
nxrB1; putative nitrate oxidoreductase subunit beta (E	330214	0	0	0	1	136	173	153	0	0	18	1
groEL; chaperonin GroEL (EC:3.6.4.9); K04077 chapero	167546	80	91	59	35	2	2	1	60	44	24	1
ccmK; carboxysome shell protein CsoS1	167555	155	162	94	38	0	0	0	54	40	39	1
putative UreA ABC transporter; substrate binding prot	167546	202	203	169	158	26	30	19	100	86	27	1
ligand-binding protein; OpuAC family	859653	7	7	8	7	51	69	74	3	1	19	1
ABC transporter	314261	17	20	19	9	30	35	36	12	15	22	1
ABC transporter; substrate-binding protein; family 5	89187	0	0	0	0	50	50	65	0	0	2	1
glnA; glutamine synthetase; glutamate--ammonia ligas	146891	62	60	54	58	3	4	2	60	53	4	1
amino acid ABC transporter substrate-binding protein	913324	3	2	2	4	33	78	43	0	0	6	1
F0F1 ATP synthase subunit beta	93058	57	63	76	34	5	4	3	39	40	47	1
peptide ABC transporter; periplasmic substrate-bindin	375451	0	2	3	0	41	48	44	0	0	6	1
glutamate/glutamine/aspartate/asparagine ABC trans	488538	2	7	11	2	31	52	44	2	3	5	1
hypothetical protein	1090946	1	1	7	0	51	47	37	0	0	5	1
ABC transporter binding protein	859653	88	68	33	29	1	4	3	38	32	10	1
rbcL; ribulose bisophosphate carboxylase; K01601 ribu	146891	46	57	40	36	1	3	0	37	41	15	1
nd	1073573	20	10	2	1	26	44	29	10	8	2	1
Scattered distribution of data (EC:3.6.1.2), KC	620202	^	^	^	^	^	^	^	^	^	^	^

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	
5	8	200	0
5	9	40	12
5	9	70	4
5	9	380	0
L	12	40	0
L	12	120	0
L	12	300	0
T	1	50	0
T	1	90	0
T	1	120	0
T	1	200	0
T	1	300	0
T	1	400	0
T	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>.)'"}
```

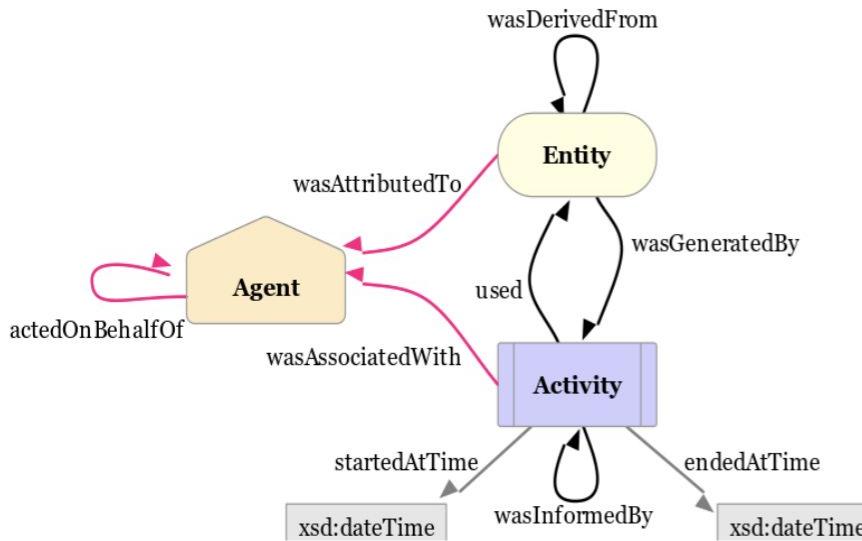
frictionlessdata.io
github.com/frictionlessdata/datapackage-pipelines



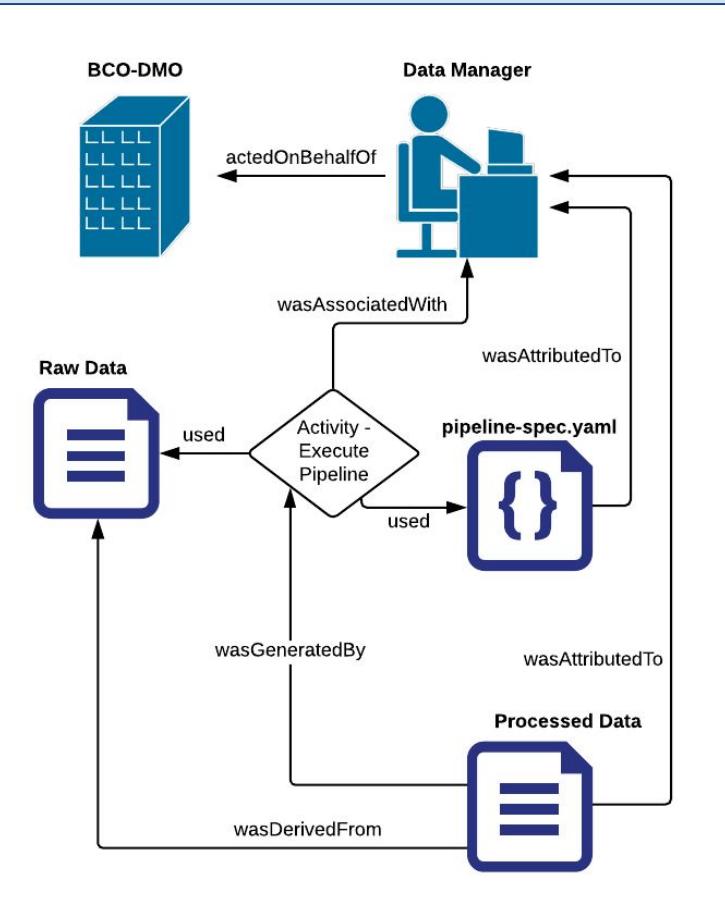
FRICTIONLESS DATA
SPECIFICATIONS AND SOFTWARE

Use Case: Provenance of Data Curation

PROV Data Model

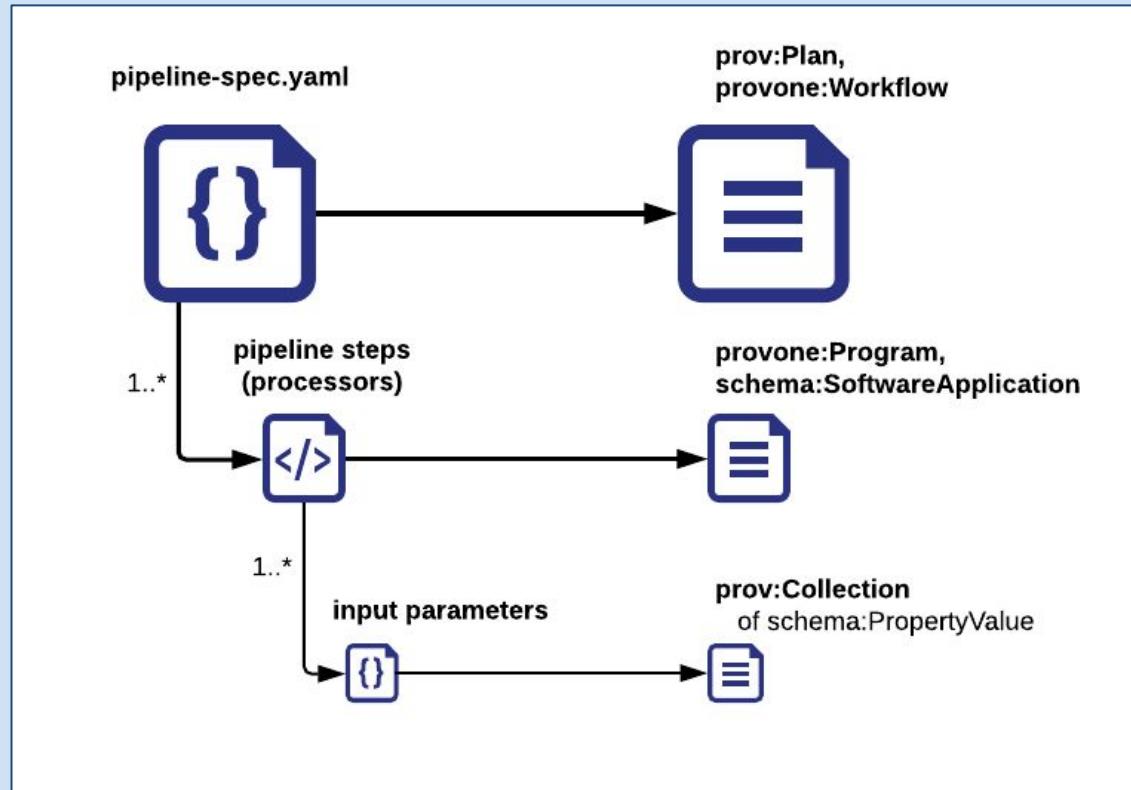


Courtesy of <https://www.w3.org/TR/prov-o/>



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

< > schema (<http://ocean-data.org/schema/>)

Active Ontology x Entities x DL Query x

Annotation properties Datatypes Individuals

Classes Object properties Data properties

Class hierarchy:

Asserted

owl:Thing
geosparql:Feature
prov:Activity
prov:Agent
prov:Entity
prov:Influence
prov:InstantaneousEvent
prov:Location
prov:Role ≡ :AgentRoleType
schema:AudioObject ≡ :AudioFile
schema:CreativeWork ≡ :CreativeWork
schema:Dataset ≡ :Dataset
schema:DigitalDocument ≡ :DigitalObject
schema:FundingAgency ≡ :FundingSource



Use Case: Provenance of Data Curation

