Towards Capturing Provenance of the Data Curation Process at Domain-specific Repositories

...

Adam Shepherd

IN53A-06

20,000 Leagues Under the Sea by Jules Verne

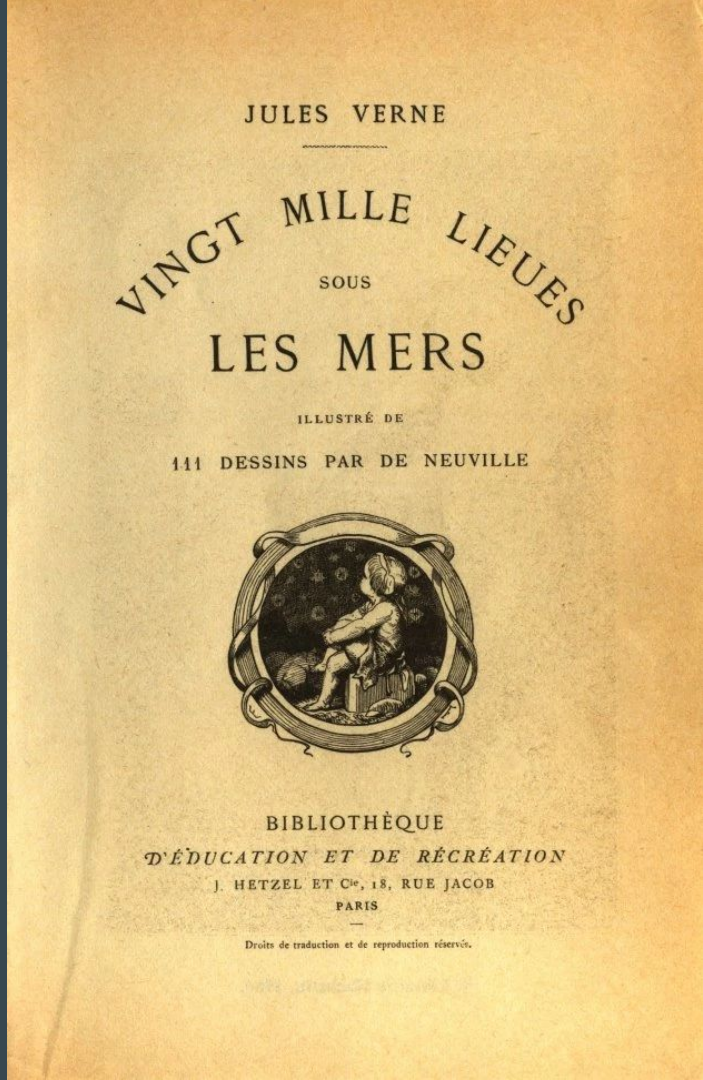
https://dentycj2qhk72.cloudfront.net/pictures/13169814171855849766.jpg

20,000 Leagues Under the Sea

Plural '**Seas**' became the singular '**Sea**'

thus...

20,000 leagues was misinterpreted to mean **depth** instead of **distance**


JULES VERNE

VINGT MILLE LIEUES
SOUS
LES MERS

ILLUSTRÉ DE
111 DESSINS PAR DE NEUVILLE

BIBLIOTHÈQUE
D'ÉDUCATION ET DE RÉCRÉATION
J. HETZEL ET Cie, 18, RUE JACOB
PARIS

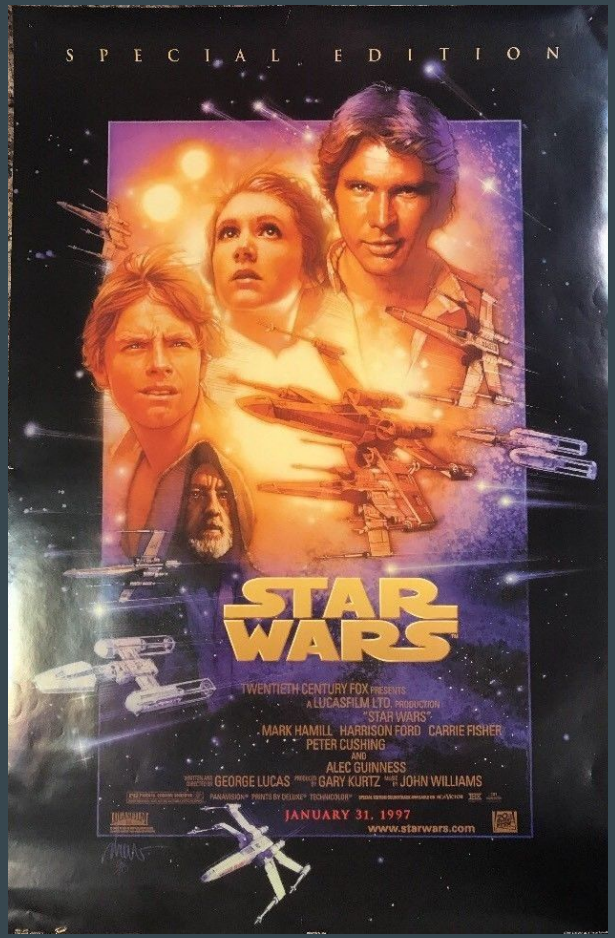
Droits de traduction et de reproduction réservés.

https://vignette.wikia.nocookie.net/childrensbooks/images/5/5e/Twenty_Thousand.jpg/revision/latest?cb=20090309165240

Lost in Translation

When variations are **perceived** to have changed the meaning of a work, there is potential for a **loss of trust**.

1977 Star Wars



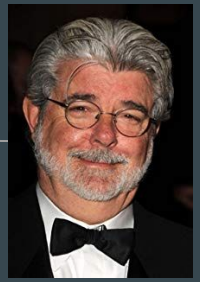
1997 Star Wars

https://cdn.shopify.com/s/files/1/0371/9417/products/Original_Star_Wars080618PF_eca1292e-66e9-417e-a811-43f2728f5542_1024x1024.jpg?v=1542461664
https://d1w8cc2yygc27j.cloudfront.net/-1472768621735153225/1072936435962636846.jpg

1997: A Disturbance in the Force

1977

1997



A single change in a split second altered the nature of a character.

Did Han Shoot First?

BCO-DMO
Biological & Chemical Oceanography Data Management Office

Submitted Columns Names

28(5,24(28))

A

st1_50m

37:2 alkenone

A Case of 'st1_50m'

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

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	
formate dehydrogenase subunit alpha (EC:1.2.1.2); K0	639282	0	0	0	0	37	41	36	0	0	3	
amino acid ABC transporter substrate-binding protein,	644966	0	0	0	0	29	38	31	0	0	1	
ligand-binding protein; OpuAC family	859653	25	16	15	9	50	55	41	12	14	19	
glutamate/glutamine/aspartate/asparagine ABC trans	488538	0	14	18	0	36	45	33	0	0	22	
nd	1073573	4	5	1	2	35	27	35	0	0	3	
amino acid ABC transporter substrate-binding protein	859653	35	31	34	16	25	24	14	25	22	26	
TonB-dependent receptor plug	518766	0	0	0	0	38	55	31	0	0	2	
hypothetical protein	926566	0	0	0	0	19	64	21	0	0	0	
chlorophyll a/b binding light harvesting protein PcbD;	146891	8	69	91	8	3	3	0	32	25	0	
Phycobilisome protein	221359	2	26	10	4	0	0	0	3	2	0	
nd	4577	22	22	16	9	9	15	9	4	5	6	
ABC transporter	439493	31	29	21	10	7	4	3	28	24	17	
TIGR00065; ftsZ; cell division protein FtsZ PF00091 20	35677	28	27	28	17	28	20	9	13	11	15	

Archive Quality

Why split these out into explicit columns if the science community prefers the original format?

Putting the FIR in FAIR

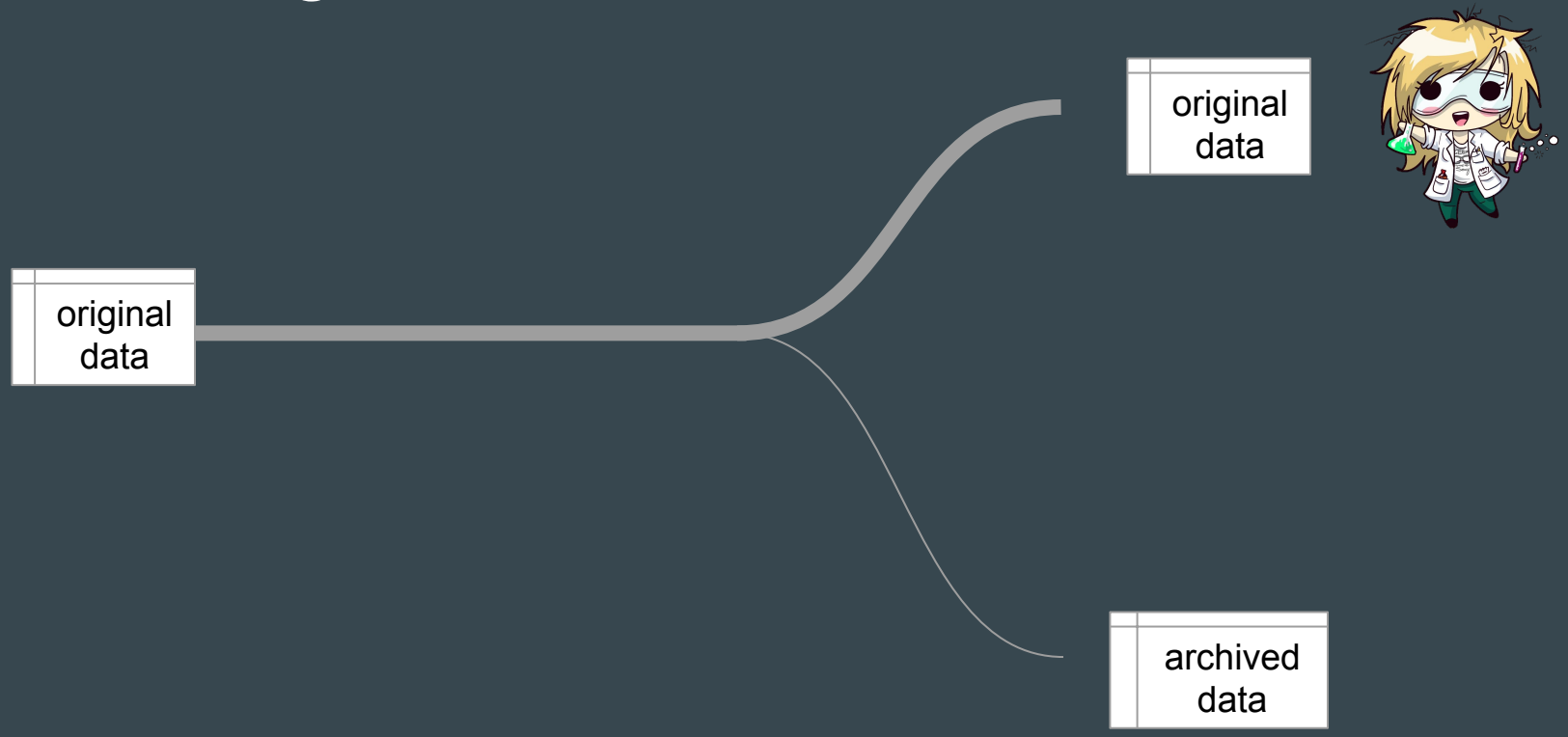
Findable - get all datasets that measured 'depth'

Interoperable - linked to community vocabs

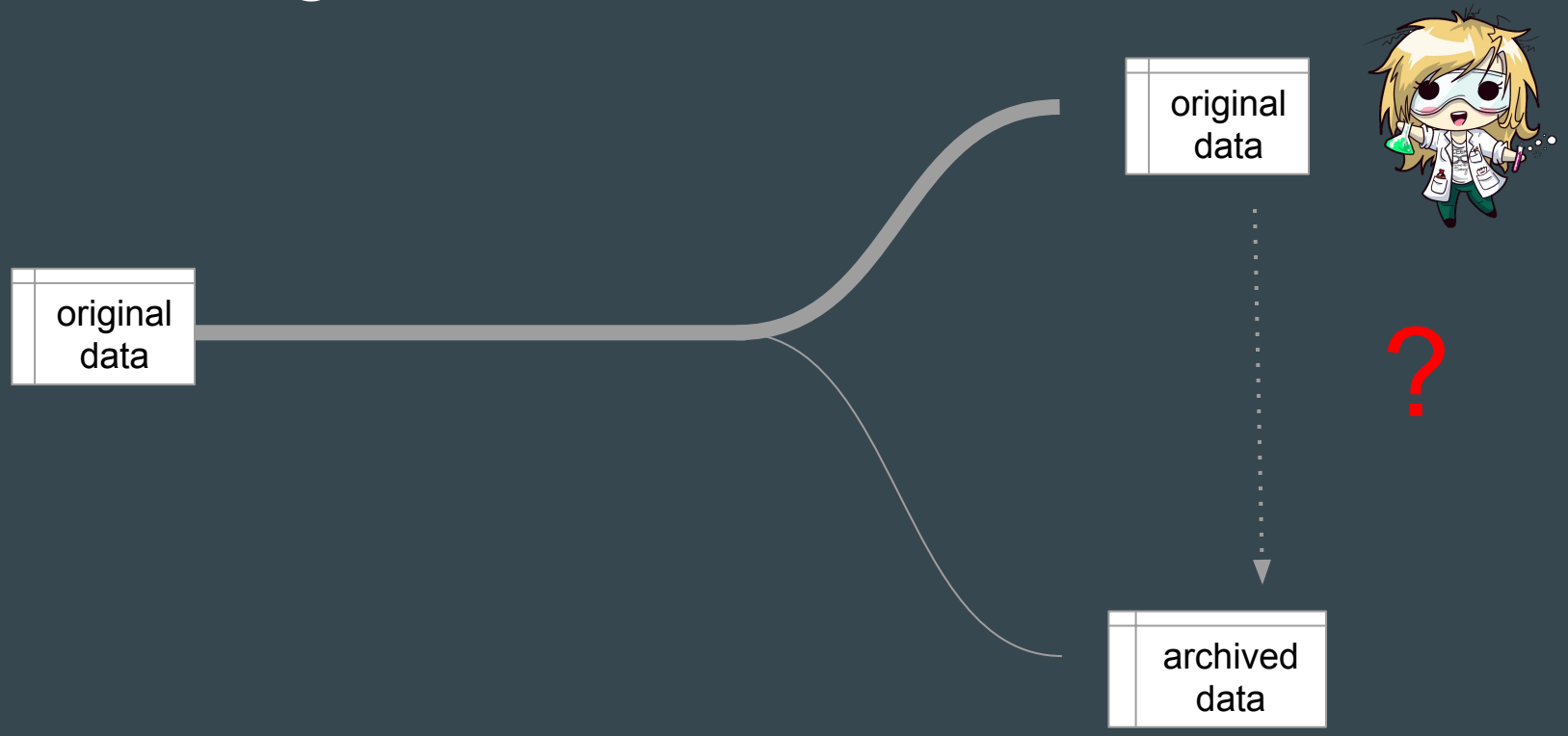
Reusable - *(meta)data with* ***provenance***

	station	depth	spectral_count
5	8	200	0
5	9	40	12
5	9	70	4
5	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
4	1	600	0
5	3	40	9
5	3	60	0

Fracturing the Community



Fracturing the Community



How do we capture provenance
of the data curation process
that links original & archived versions
in a reusable way ?

Declarative Workflows

```
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: add_resource
    parameters:
        name: mcmurdo_epifauna,
        url: 'http://datadocs.bco-dmo.org/docs/TestProject/data_docs/latlon_DDM_to_DD/McMurdoEpifauna.xlsx',
        format: xlsx,
        sheet: animals,
        headers: 1,
```

Declarative Workflows - A set of steps to execute

```
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: add_resource
    parameters:
        name: mcmurdo_epifauna,
        url: 'http://datadocs.bco-dmo.org/docs/TestProject/data_docs/latlon_DDM_to_DD/McMurdoEpifauna.xlsx',
        format: xlsx,
        sheet: animals,
        headers: 1,
```

Declarative Workflows - Each step is "named"

```
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: add_resource
    parameters:
        name: mcmurdo_epifauna,
        url: 'http://datadocs.bco-dmo.org/docs/TestProject/data_docs/latlon_DDM_to_DD/McMurdoEpifauna.xlsx',
        format: xlsx,
        sheet: animals,
        headers: 1,
```

Declarative Workflows - Each step has inputs

```
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
   \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
   \ 40.1'E\"). "
  pipeline:
  - run: add_resource
    parameters:
        name: mcmurdo_epifauna,
        url: 'http://datadocs.bco-dmo.org/docs/TestProject/data_docs/latlon_DDM_to_DD/McMurdoEpifauna.xlsx',
        format: xlsx,
        sheet: animals,
        headers: 1,
```

Declarative Workflows - More steps

```
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: ad
    paramet
        nam
        url
        for
        she
        hea
```

```
  - 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>.)"}
```

Declarative Workflows - Names identify code to execute

```yaml
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: ad
    paramet
        nam
        url
        for
        she
        hea
```

```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>.)"}
```

Declarative Workflows - Each step has its own inputs

```yaml
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: ad
    paramet
        nam
        url
        for
        she
        hea
```

```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>.)"}
```

Benefits of Declarative Workflows

Pipeline Title: lat_lon_DDM_to_DD

Pipeline Description: Add lat and lon columns in decimal degrees (DD) given one column with lat and lon in format degrees decimal minutes (DDM) (e.g. "77° 51.3'S 166° 40.1'E").

1

Add resource

Processor: Add resource

Name must match the `^([-a-z0-9._/])+$` regular expression.

Name: mcmurdo_epifauna

URL: http://datadocs.bco-dmo.org/docs/TestProject/data_docs

Format: xlsx

Header row #: 1

Because declarative workflows
build a configuration (WHAT)
instead of code (HOW)

Build tools to:

- construct these configurations,
- execute the workflow,
- convert the configuration into provenance

Benefits of Declarative Workflows

Pipeline Title: lat_lon_DDM_to_DD

Pipeline Description: Add lat and lon columns in decimal degrees (DD) given one column with lat and lon in format degrees decimal minutes (DDM) (e.g. "77° 51.3'S 166° 40.1'E").

1

Add resource

Processor: Add resource

Name must match the `^([-a-z0-9._/])+$` regular expression.

Name: mcmurdo_epifauna

URL: http://datadocs.bco-dmo.org/docs/TestProject/data_docs

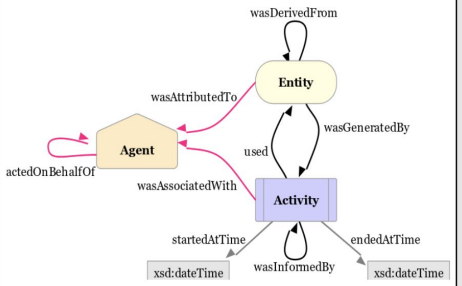
Format: xlsx

Header row #: 1

```yaml
lat_lon_DDM_to_DD:
  title: lat_lon_DDM_to_DD
  description: "Add lat and lon columns in decimal degrees (DD) given one column with\
    \ lat and lon in format degrees decimal minutes (DDM) (e.g. \"77\xB0 51.3'S 166\xB0\
    \ 40.1'E\"). "
  pipeline:
  - run: add_resource
    parameters:
        name: mcmurdo_epifauna,
        url: 'http://datadocs.bco-dmo.org/docs/TestProject/data_docs/latlon_DDM_to_DD/McMurdoEpifauna.xlsx',
        format: xlsx,
        sheet: animals,
        headers: 1,
  - run: stream_remote_resources
    cache: True
    parameters:
      resources: [mcmurdo_epifauna]
  - run: set_types
    cache: True
    parameters:
      resources: [mcmurdo_epifauna]
      types:
        year: {type: number}
        site: {type: string}
        lat_long: {type: string}
        genus_species: {type: number}
  - run: bcodmo_pipeline_processors.add_schema_metadata
    cache: True
    parameters:
      resources: [mcmurdo_epifauna]
      missingValues: ["nd"]
  - 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>.)"}
  - run: bcodmo_pipeline_processors.round_fields
    cache: True
    parameters:
      resources: [mcmurdo_epifauna]
      fields:
      - {digits: 5, name: lat_converted}
  - run: bcodmo_pipeline_processors.round_fields
    cache: True
```

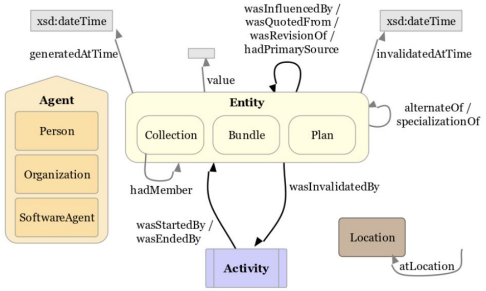
From Declarative Workflows to Provenance

PROV Data Model

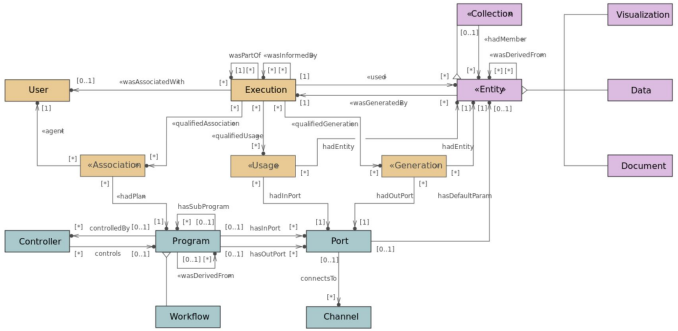


Courtesy of https://www.w3.org/TR/prov-o/

PROV-O Extended Data Model

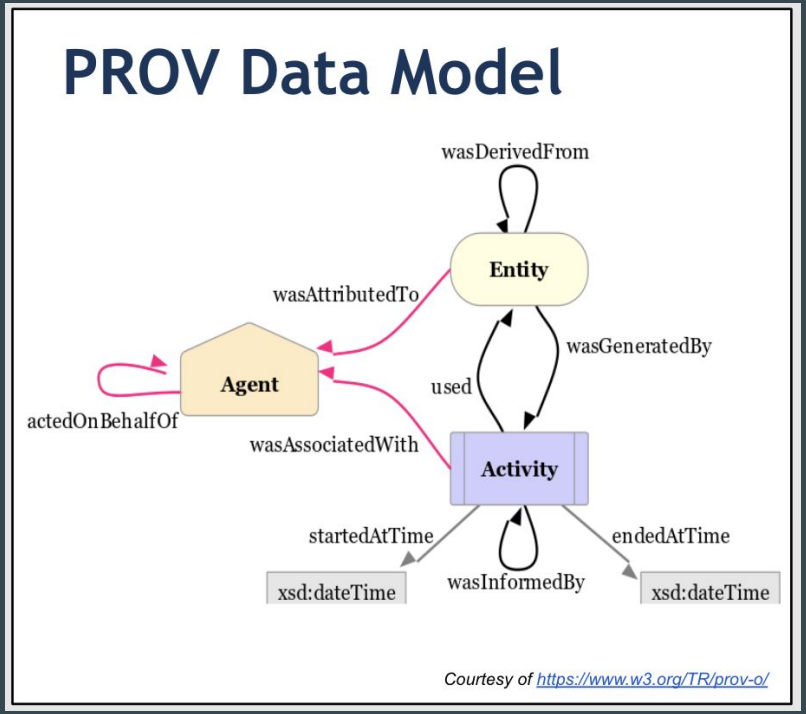


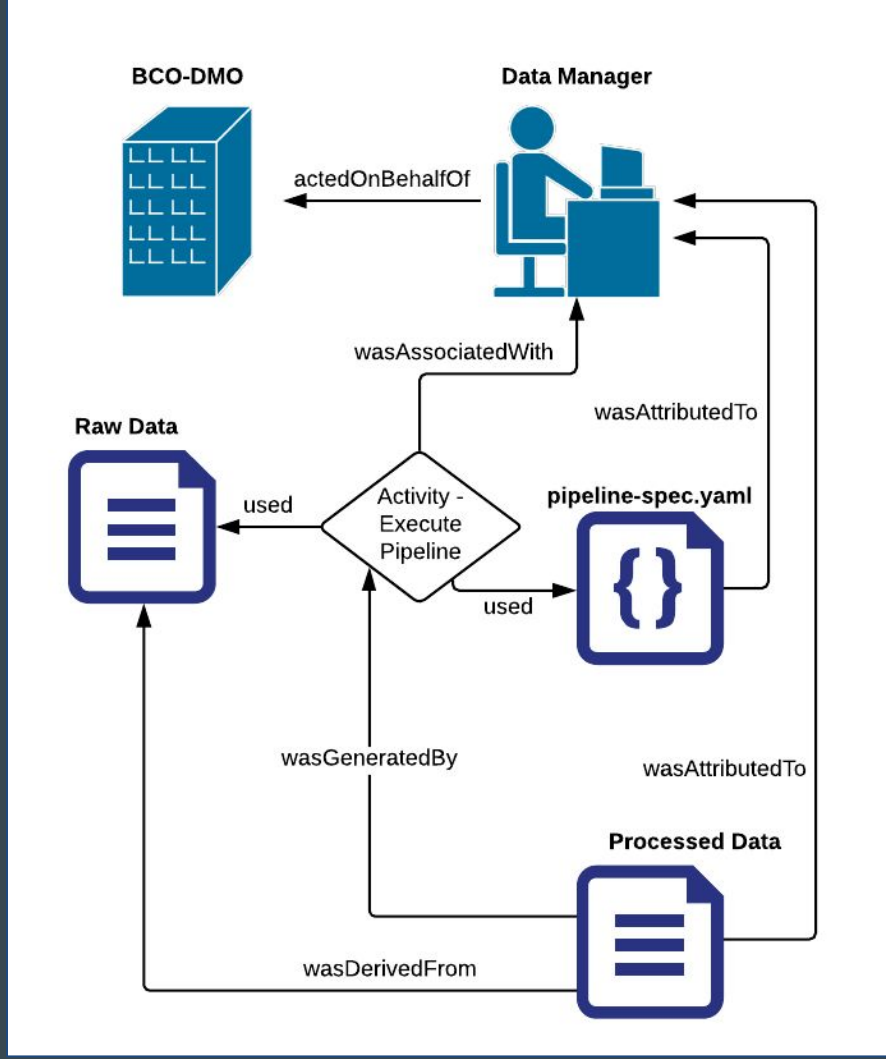
PROVONE Data Model



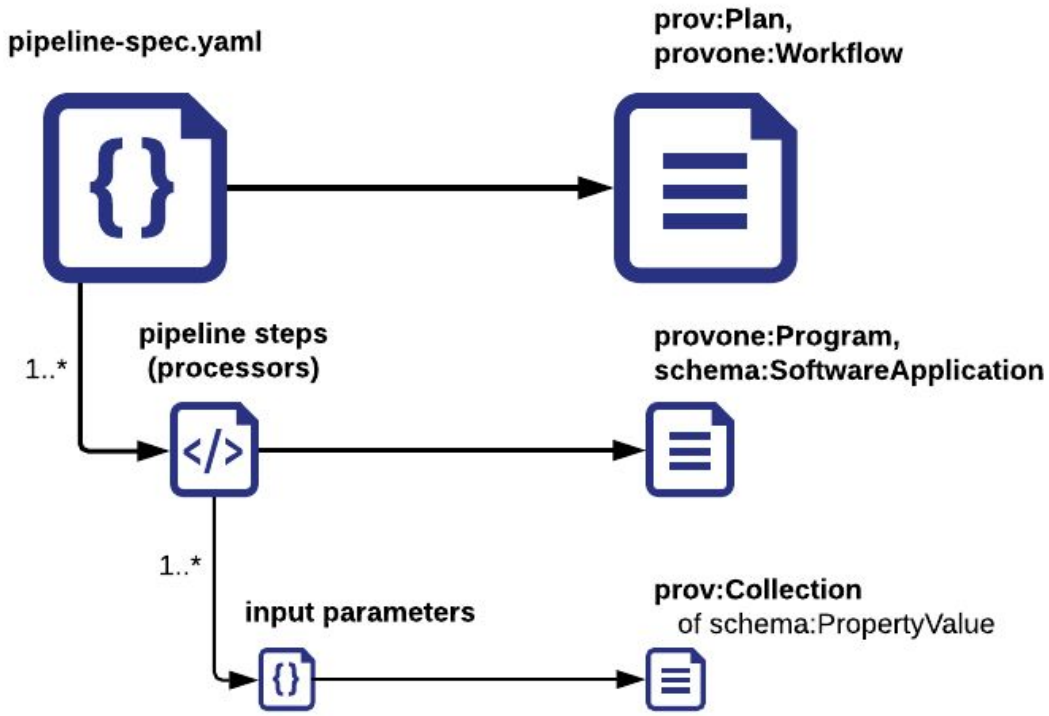
Courtesy of https://purl.dataone.org/provone-v1-dev

Concept Map in PROV Data Model

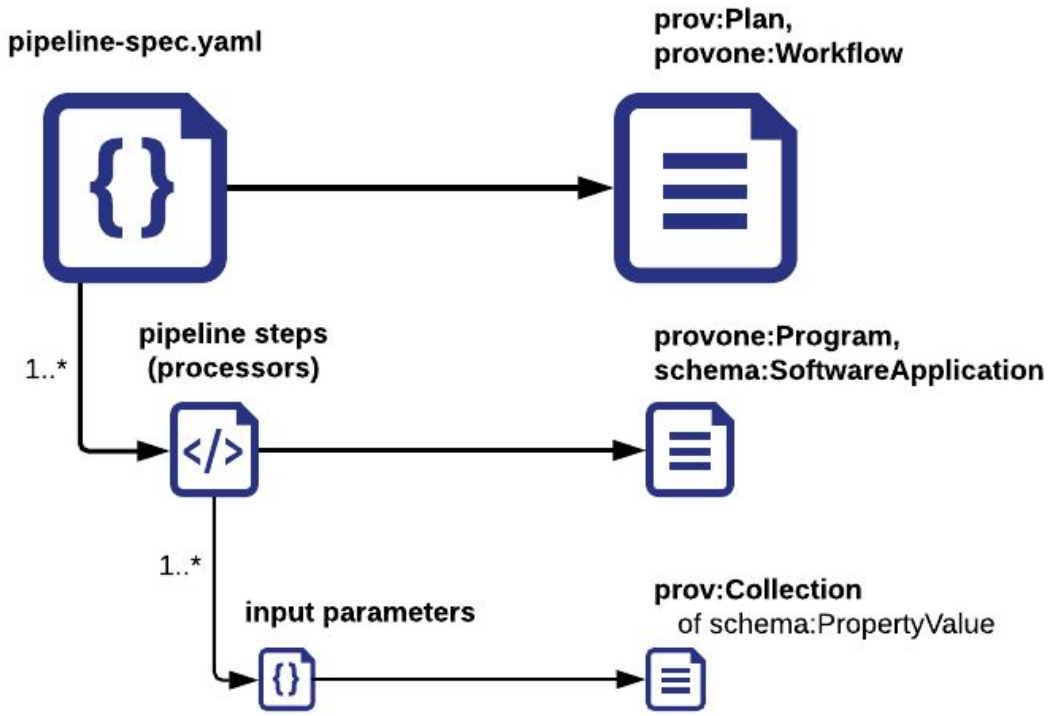




PROV-O + Schema.org



PROV-O + Schema.org




```
Sub propertyValueForParameter(parameter)
  prop_value = new PropertyValue(parameter.name)

  If parameter.value  is an Object
      prop_value = propertyValueForParameter(parameter.value)
  Else
    prop_value = parameter.value

  Return prop_value
End Sub
```

Workflow to PROV

```rdf
@prefix : <http://data.example.org/id/dataset/1234/v1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix provone: <http://purl.dataone.org/provone/2015/01/15/ontology#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

: a prov:Bundle,
        prov:Entity ;
    prov:generatedAtTime "2018-09-21T13:38:10+00:00"^^xsd:dateTime ;
    prov:wasAttributedTo :alice .

:frictionless-data-pkg a schema:DigitalDocument,
        prov:Data,
        prov:Entity ;
    schema:encodingFormat "application.vnd.datapackage+json"^^xsd:string ;
    schema:url "https://example.org/dataset/1234/v1/datapackage.json"^^xsd:anyURI ;
    prov:qualifiedGeneration [ a prov:Generation ;
            prov:activity :executed-pipeline ;
            prov:endTime "2018-09-21T13:38:10+00:00"^^xsd:dateTime ;
            prov:startTime "2018-09-21T13:37:53+00:00"^^xsd:dateTime ] ;
    prov:wasGeneratedBy :executed-pipeline .

:processed-data a schema:Dataset,
        prov:Entity ;
    schema:distribution [ a schema:DataDownload ;
            schema:contentUrl "https://example.org/dataset/1234/v1/McMurdoEpifauna.csv"^^xsd:anyURI ;
            schema:encodingFormat "text/csv"^^xsd:string ] ;
    prov:hadPrimarySource :raw-data ;
    prov:qualifiedGeneration [ a prov:Generation ;
            prov:activity :executed-pipeline ;
            prov:endTime "2018-09-21T13:38:09+00:00"^^xsd:dateTime ;
            prov:startTime "2018-09-21T13:37:54+00:00"^^xsd:dateTime ] ;
    prov:wasDerivedFrom :pipeline-spec ;
    prov:wasGeneratedBy :executed-pipeline .

:step-1-add-resource a provone:Program,
        prov:Entity ;
    schema:supportingData :step-1-add-resource-inputs .

:step-1-add-resource-inputs a schema:DataFeed ;
    schema:dataFeedElement [ a prov:Collection ;
            rdfs:comment "A single step in pipeline."@en-US ;
            prov:hadMember [ a provone:Data,
                        schema:PropertyValue,
                        prov:Entity ;
                    schema:name "run"^^xsd:string ;
                    schema:value "add_resource"^^xsd:string ],
                [ a provone:Data,
                        schema:PropertyValue,
                        prov:Entity ;
                    schema:name "parameters"^^xsd:string ;
                    schema:value [ a schema:PropertyValue ;
                            schema:name "headers"^^xsd:string ;
                            schema:value 1 ],
                        [ a schema:PropertyValue ;
```

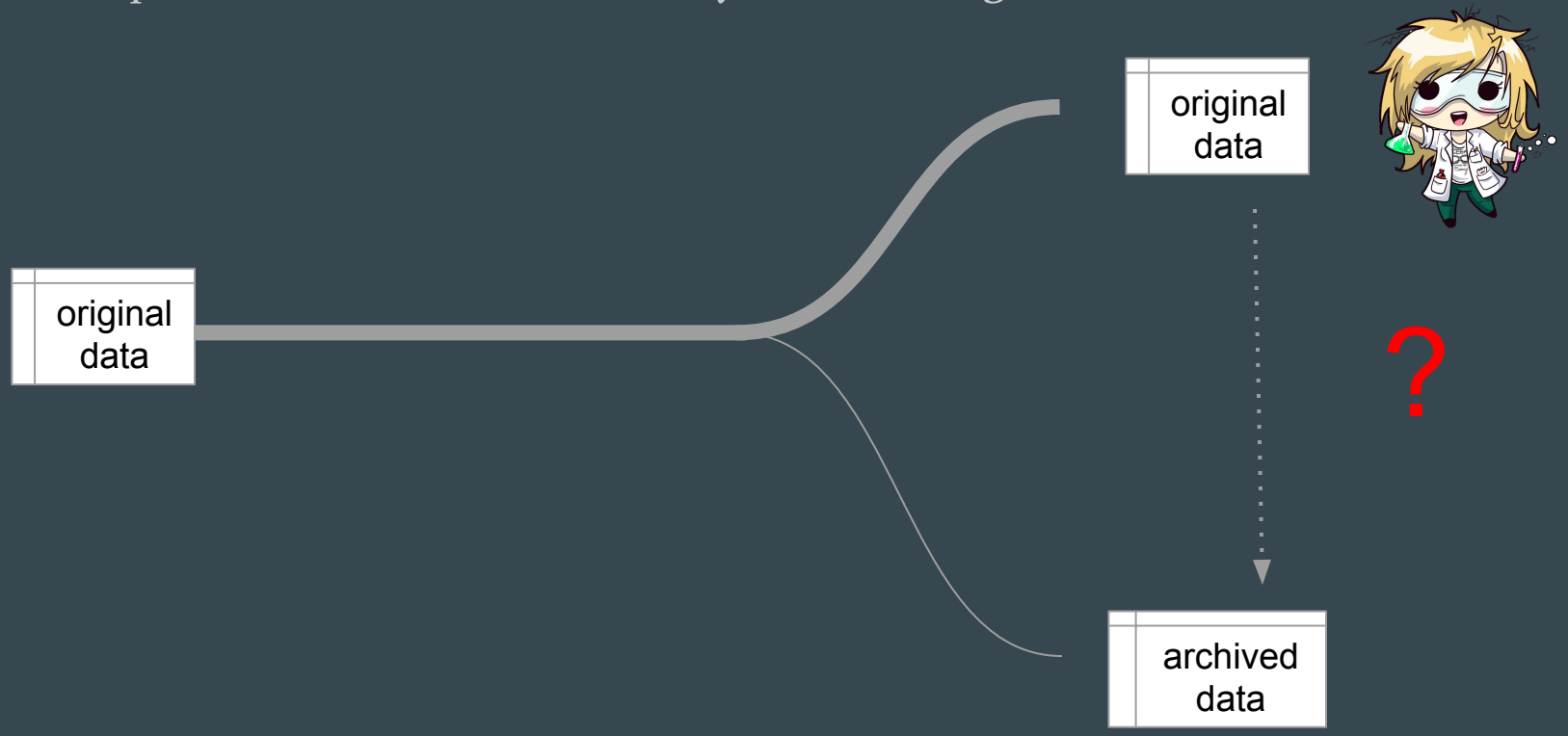
```rdf
:step-1-add-resource-inputs a schema:DataFeed ;
    schema:dataFeedElement [ a prov:Collection ;
            rdfs:comment "A single step in pipeline."@en-US ;
            prov:hadMember [ a provone:Data,
                        schema:PropertyValue,
                        prov:Entity ;
                    schema:name "run"^^xsd:string ;
                    schema:value "add_resource"^^xsd:string ],
                [ a provone:Data,
                        schema:PropertyValue,
                        prov:Entity ;
                    schema:name "parameters"^^xsd:string ;
                    schema:value [ a schema:PropertyValue ;
                            schema:name "headers"^^xsd:string ;
                            schema:value 1 ],
                        [ a schema:PropertyValue ;
                            schema:name "name"^^xsd:string ;
                            schema:value "mcmurdo_epifauna"^^xsd:string ],
                        [ a schema:PropertyValue ;
                            schema:name "url"^^xsd:string ;
                            schema:value "https://example.org/dataset/1234/original/20180921T123456Z/McMurdoEpifauna.xlsx"^^x
                        [ a schema:PropertyValue ;
                            schema:name "format"^^xsd:string ;
                            schema:value "xlsx"^^xsd:string ],
                        [ a schema:PropertyValue ;
                            schema:name "sheet"^^xsd:string ;
                            schema:value "animals"^^xsd:string ] ] ] .
```

What use is this PROV?

The provenance is data that directly links the original data to the archived version

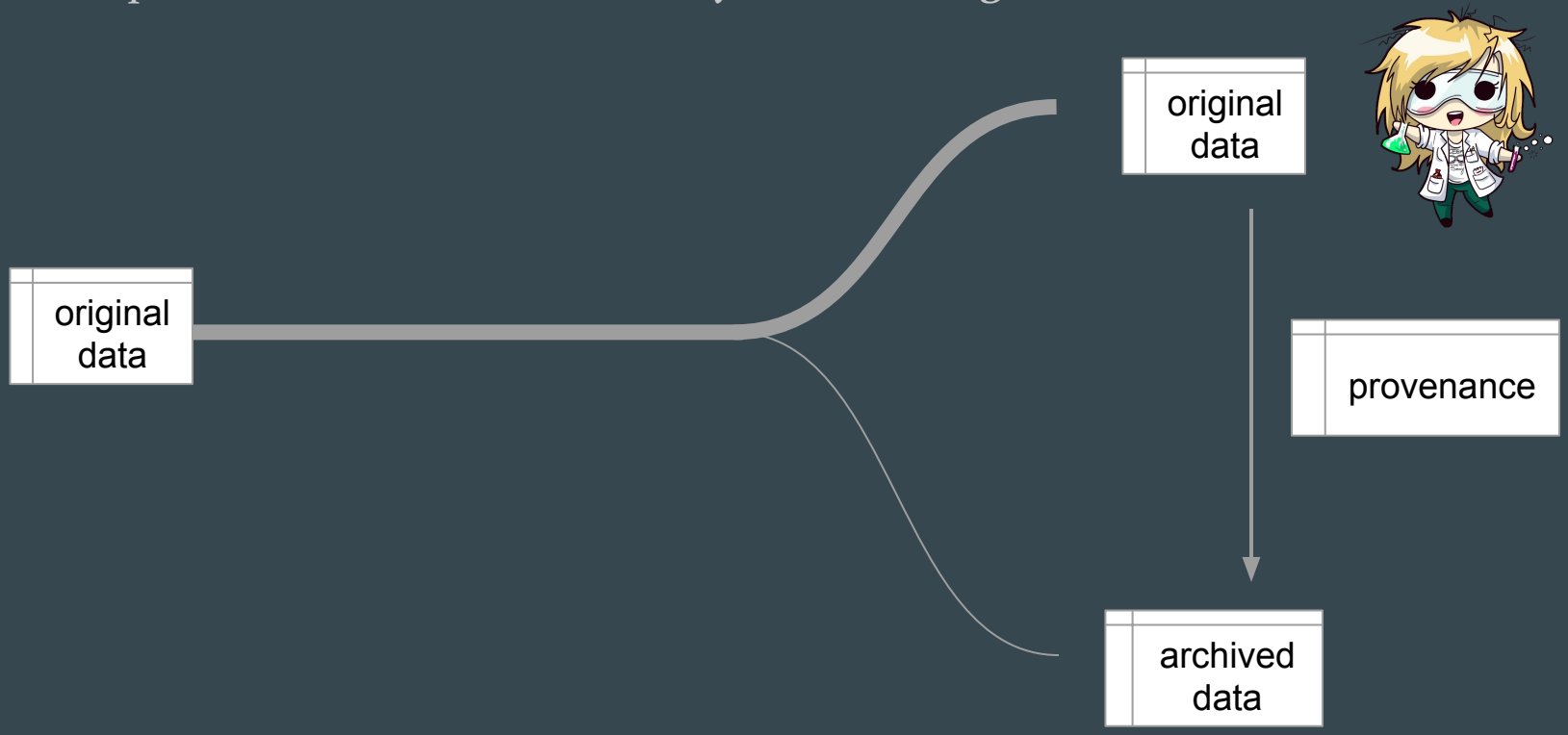
What use is this PROV?

The provenance is data that directly links the original data to the archived version



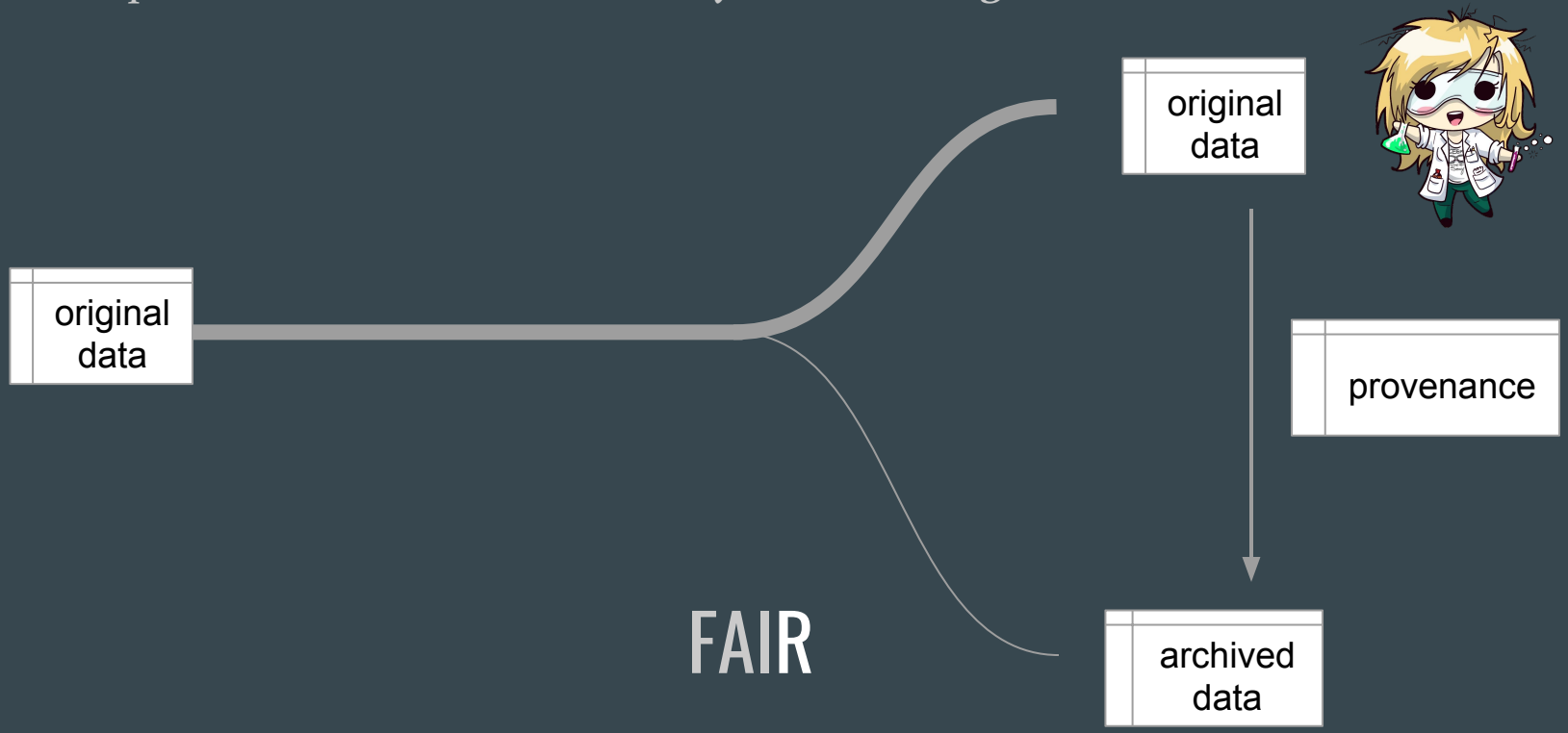
What use is this PROV?

The provenance is data that directly links the original data to the archived version



What use is this PROV?

The provenance is data that directly links the original data to the archived version



Workflow Tools at BCO-DMO

frictionlessdata.io

github.com/frictionlessdata/datapackage-pipelines

pypi.org/project/dataflows

w3.org/TR/prov-o/

purl.dataone.org/provone-v1-dev

schema.org

Questions?