What are we DOIng about Dynamic Data at Ocean Networks Canada?

Portage Network Webinar 2020-11-17

Ocean Networks Canada, ROR: 05qknh003
University of Victoria, ROR: 04s5mat29

Presenters:

- **Reyna Jenkyns, ORCID: 0000-0001-6975-6816**
- **Melissa Cuthill, ORCID: 0000-0002-3987-865X**
- **Chantel Ridsdale, ORCID: 0000-0001-7218-1232**

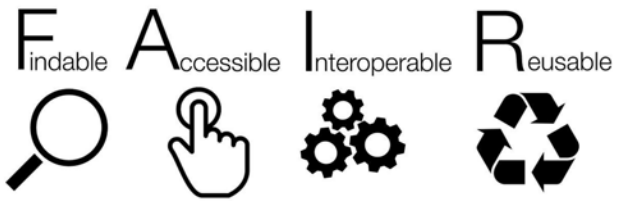
A UNIVERSITY OF VICTORIA INITIATIVE

Ocean Networks Canada

- ONC data are very **dynamic** due to continually accumulating data streams, data reprocessing and data product code versioning
- Highly **heterogeneous** – fixed and mobile platforms, instrument types, data formats and processing levels, real-time vs autonomous

Setting the Stage

R13: Data Discovery and Identification - The repository enables users to discover the data and refer to them in a persistent way through proper citation.



COPDESS (Coalition for Publishing Data in the Earth and Space Sciences) - Commitment Statement includes an emphasis on data citation. ONC is a signatory.

2018 - Research Data Management funding program introduced following community consultation with an emphasis on FAIR.
9 projects awarded funding ($3.2M total)

MINTED = Making Identifiers Necessary for Tracking Evolving Data

RDA Data Citations WG Guidelines

Research Data Alliance:

- over 10,000 members from 145 countries
- a neutral space for members to develop & adopt infrastructure that promotes data-sharing & data-driven research

RDA Data Citations WG Guidelines: Rauber, A., et al, Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use (2016)

- ***Abstract Excerpt:*** Research **data is changing over time** as new records are added, errors are corrected and obsolete records are deleted from a data set. Scholars rarely use an entire data set or stream data as it is, but rather select **specific subsets** tailored to their research questions. In order to keep such experiments reproducible and to share and cite the particular data used in a study, researchers need means of **identifying the exact version of a subset as it was used** during a specific execution of a workflow, even if the data source is continuously evolving. ..we present **14 recommendations** on how to adapt a data source for providing identifiable subsets for the long term, elaborated by the RDA Working Group on Dynamic Data Citation (WGDC). The proposed solution is based upon **versioned data, timestamping and a query based subsetting** mechanism.

- R1 - Data Versioning, R2 - Timestamping, R3 - Query Store Facilities, R4 - Query Uniqueness, R5 - Stable Sorting, R6 - Result Set Verification, R7 - Query Timestamping, R8 - Query PID, R9 - Store the Query, R10 - Automated CitationTexts, R11 - Landing Page, R12 - Machine Actionability, R13 - Technology Migration, R14 - Migration Verification

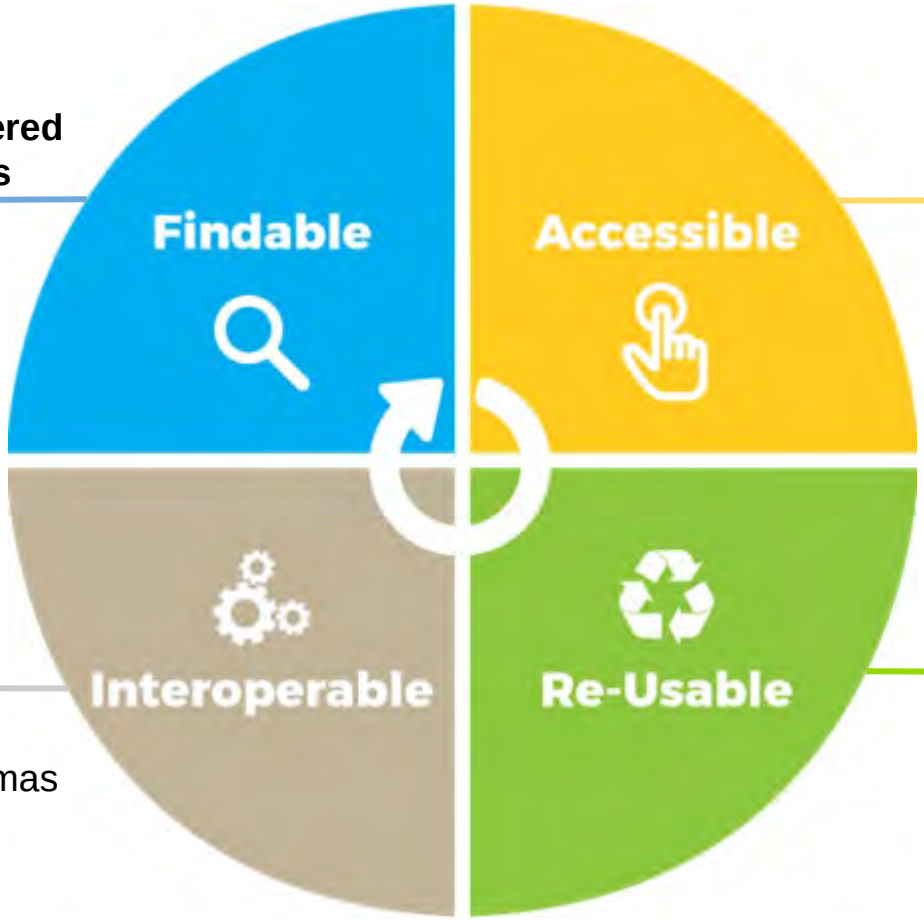
The FAIR Principles

Guidelines for data publishers to improve data **discovery** and **reuse**

- Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18

Data can be discovered by interested parties

- PIDs
- rich metadata
- searchable



Data is retrievable using its PID

Or, if the data is gone, the PID still links to the metadata that tells you what happened to it.

Data & metadata are in standardized formats

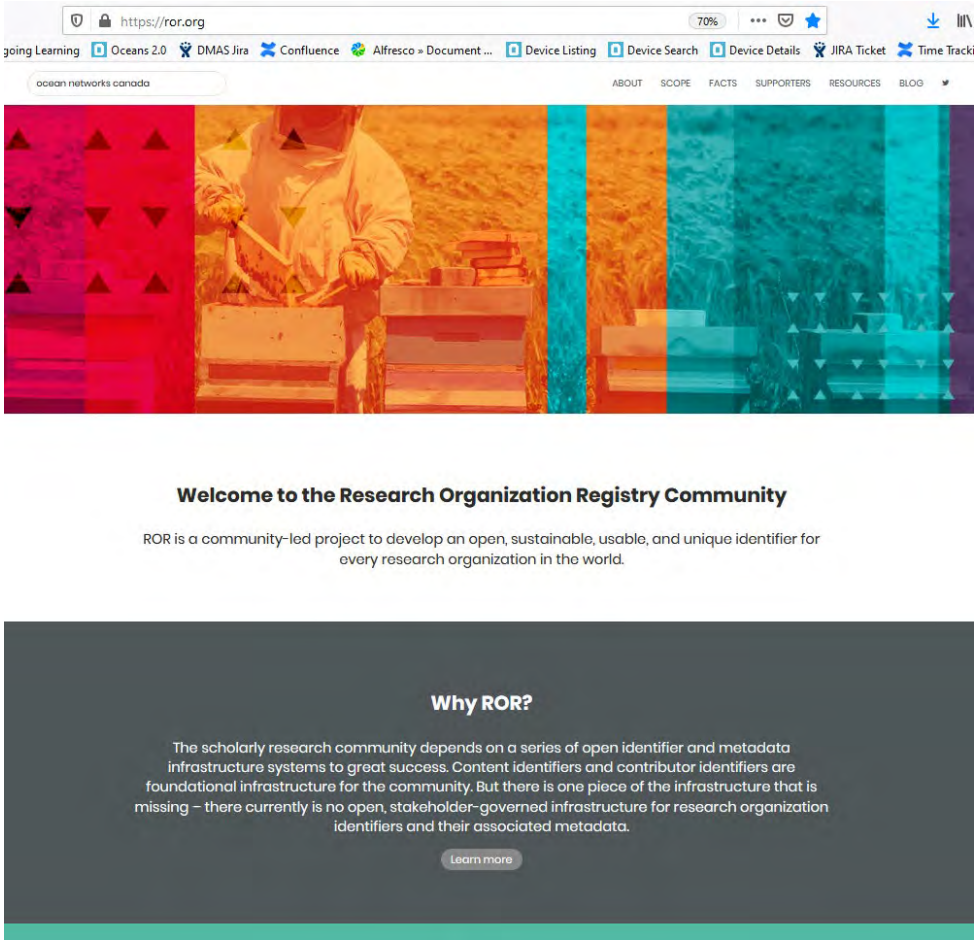
- file formats
- metadata schemas
- controlled vocabularies

Data is optimized for re-use

- license
- provenance
- community standards

Research Organization Registry

- Persistent identifiers for research organizations
- Launched in January 2019
- ~99,000 organizations as of November 2020



TRUST Principles

https://www.nature.com/articles/s41597-020-0486-7.pdf

Transparency	To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
Responsibility	To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
User Focus	To ensure that the data management norms and expectations of target user communities are met.
Sustainability	To sustain services and preserve data holdings for the long-term.
Technology	To provide infrastructure and capabilities to support secure, persistent, and reliable services.

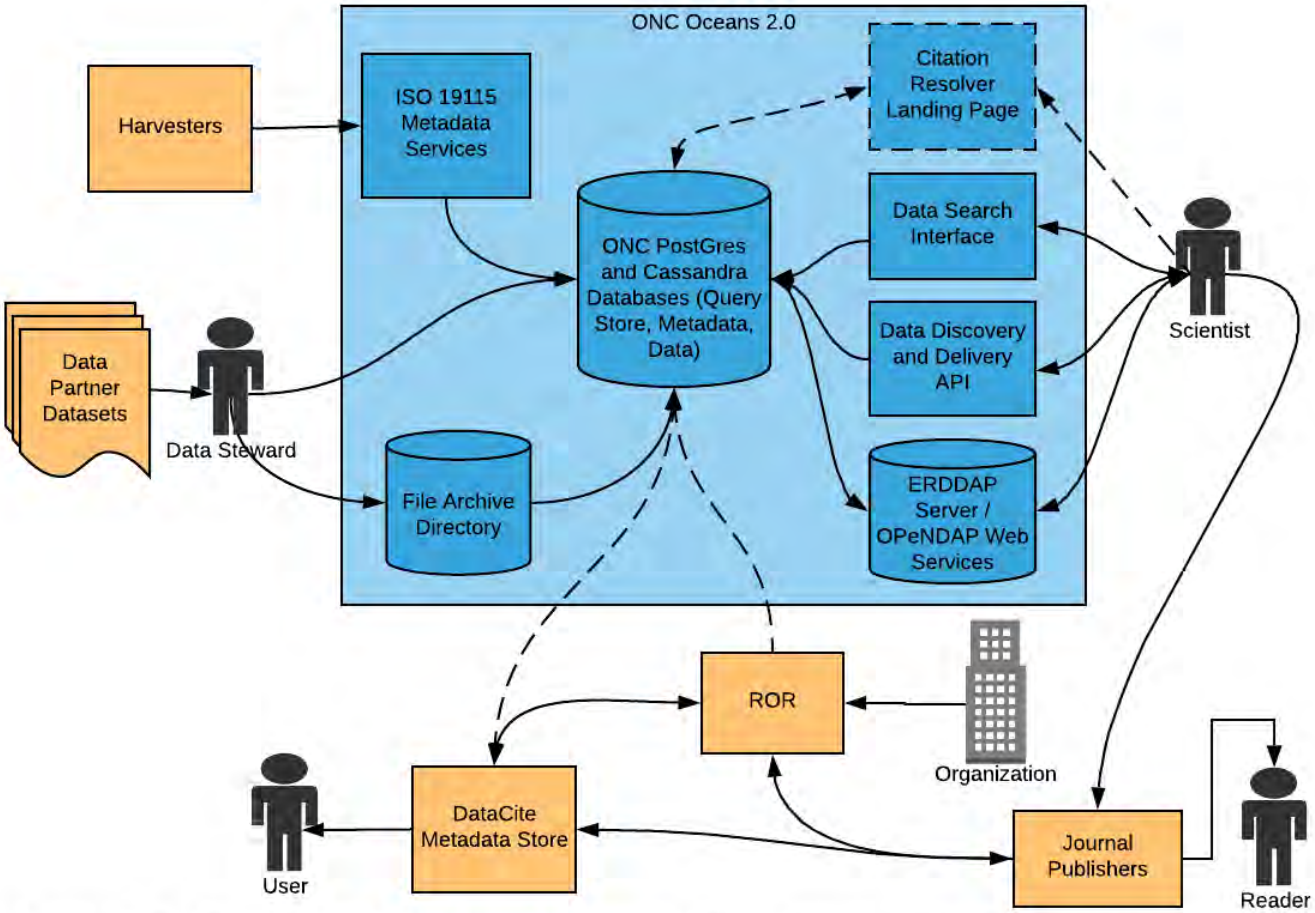
".. to make data **FAIR** whilst preserving them over time requires **trustworthy digital repositories** (TDRs) with **sustainable governance** and **organizational frameworks, reliable infrastructure,** and **comprehensive policies supporting community-agreed practices**. TDRs, with their clear remit to actively **preserve data** in response to changes in both technology and stakeholder requirements, play an important role in **maintaining the value of data**. They are held in a position of trust by their users as they **accept the responsibilities of data stewardship**."

Lin, D., Crabtree, J., Dillo, I., Jenkyns,R., *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

MINTED

- Implement **dynamic data citations**, applying 14 recommendations from RDA Data Citations WG output which covers versioning, query store, resolver landing page, technology migration resilience
- Improve **provenance, versioning, and ISO 19115 metadata records** as they relate to data citation framework
- Utilize **DataCite** services for registering datasets
- Introduce RORs for organizational dataset contributors and user accounts, leveraging ROR frameworks and advice
- Deliver citation text provision service and a citation resolver services to **National Data Services Framework**
- **Design challenge** given the permanent nature of DOIs – important decisions for **dataset granularity** and **sustainable architecture**
- **Cultural challenge** with agile and innovative culture

System Architecture



System architecture description: The ONC Oceans 2.0 system (in blue), and third party sources and applications (in orange). Dotted lines indicate aspects that need to be added, while all ONC components would be modified. The ONC components can be directly controlled via the project, with expected modifications to include a new data model and tables within the database, additional web services, integration of third party APIs, and data citation features.

Data Granularity

- Challenges in dataset granularity **boundary** decisions
  - By time?
  - By geography?
  - By instrument type?
  - By platform?
  - By data product level?
- **Constraints** to consider from DataCite metadata kernel, RDA guidelines, suitability to ONC data architecture, data partner attributions, end-users
- **RDA Data Granularity WG**
  - compatible community conventions are important for interoperability networks of repositories, but many diverse existing approaches and terminology interpretations; granularity decisions also impact discovery, access, citations and more.
  - BoF at RDA Plenary 16, case statement to be submitted soon

What is a dataset at ONC?

1 Dataset = 1 Deployment of 1 Device

i.e. Device A at Site B, from Date X to Date Y

From this...

...to this.

How and when to initially mint a dataset?

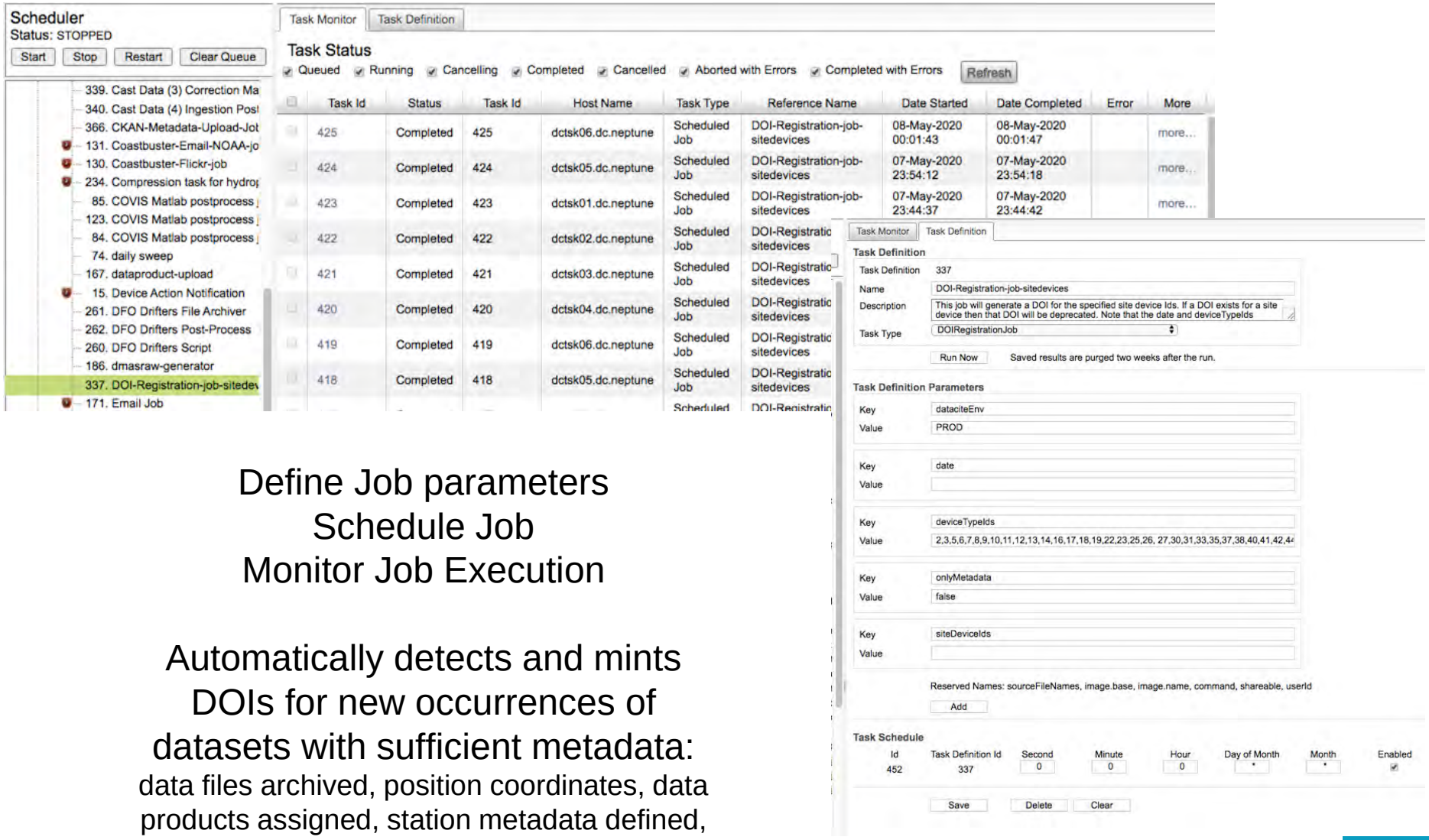
QUESTIONS:

- How soon after an instrument is deployed and data is streaming in data?
- How to automate ensuring all required metadata is in database and deemed correct?
- Manual or automated integration into our instrument workflows?
- Wait until a query exists in the query store?
- Mass or batch minting of existing datasets in repository?
- Data steward peer reviews and quality assessments are complete?

DECISION – Daily scheduled job created to regularly check if there is a new deployment for any devices and register DOI if all ”minimum” metadata is populated

- minimum is a combination of DataCite requirements plus fields determined necessary at ONC, changes will be accounted through versioning.

DataCite DOI Minting - *Automation*



Define Job parameters
Schedule Job
Monitor Job Execution

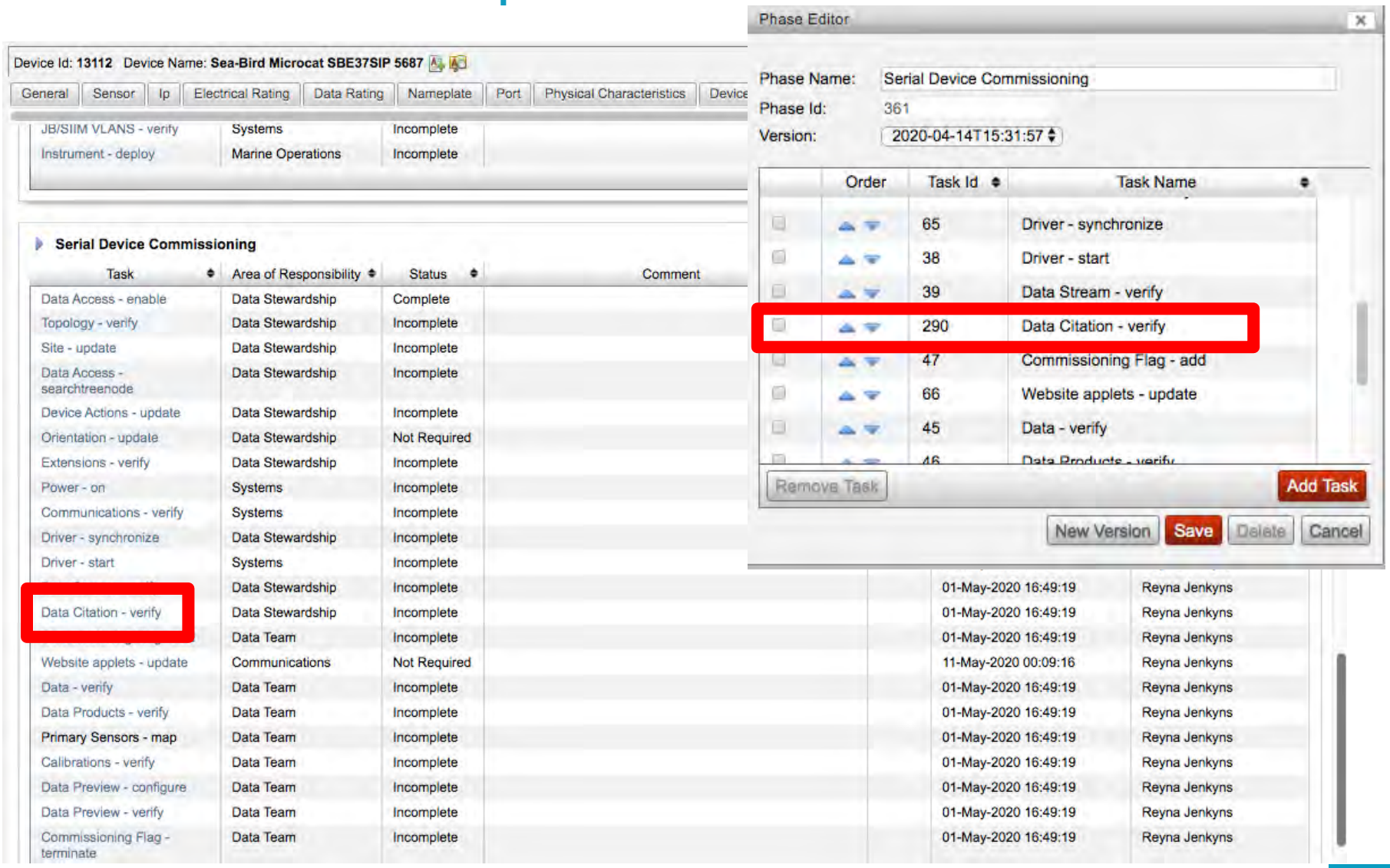
Automatically detects and mints DOIs for new occurrences of datasets with sufficient metadata:
data files archived, position coordinates, data products assigned, station metadata defined, organizational metadata defined

Automated Abstract Deconstructed

Construction: The DeviceName was deployed on SiteDeviceDateFrom *at/on* SearchTreeNodeName. SearchTreeNodeDescription. This device is a DeviceCategoryName. DeviceCategoryDescription. It was deployed on a Fixed/Mobile/Profiling platform. Data from this deployment were archived and made available through Ocean Network Canada's Oceans 2.0 digital infrastructure, with quality assurance and derived data products following established practices.

Example: The WET Labs ECO FLNTUS 4670 was deployed on 2019-05-16 at Upper Slope. Upper Slope is a location within Barkley Canyon, which is located on the upper continental slope. This device is a Fluorometer Turbidity. Fluorometer Turbidity instruments measure chlorophyll fluorescence and turbidity within the same volume of seawater. The instrument uses a light emitting diode (LED) to provide an excitation source. The fluoresced light is received by a detector at a particular angle from the LED source, and uses an interference filter to discriminate against scattered excitation light. Turbidity is measured at the same time, by detecting scattered light from a LED, which is positioned at the same angle as the chlorophyll fluorescence. It was deployed on a fixed platform. Data from this deployment were archived and made available through Ocean Network Canada's Oceans 2.0 digital infrastructure, with quality assurance and derived data products following established practices.

Data Stewardship Verification - Device Workflows



https://data.oceannetworks.ca/DeviceListing?DeviceId=13112&Tab=workflow

Benefits of PIDs for Datasets

★ Using PIDS makes finding and citing data EASY

Compare with our previous approach to data citation...

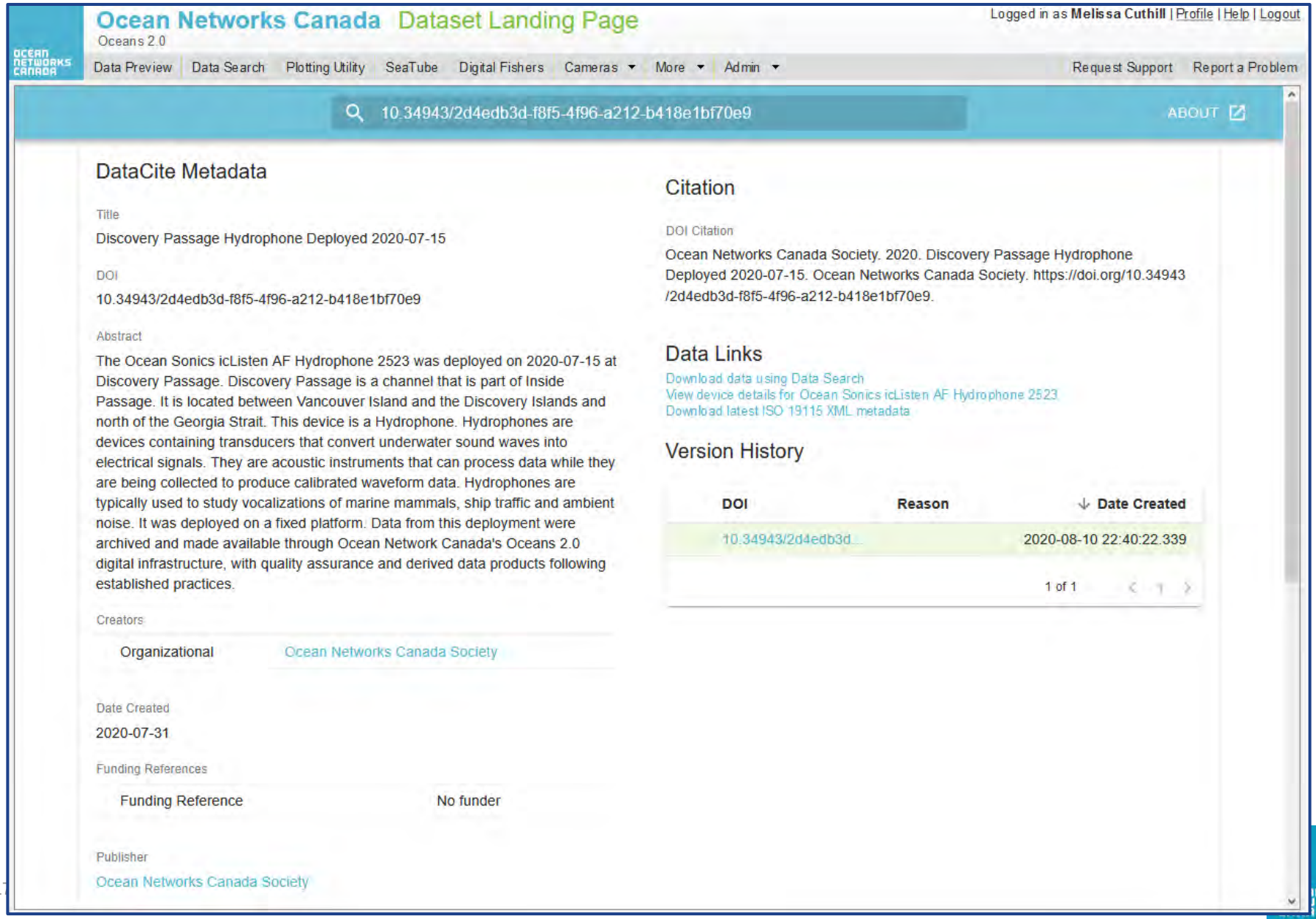
THEN

"Should publications include the option of a data **citation**, please consult the metadata information provided for the exact citation(s) associated with the data received from Ocean Networks Canada. An **example** citation might read:
Ocean Networks Canada Data Archive, http://www.oceannetworks.ca, bottom pressure data from Clayoquot Slope from 25 January 2010 to 25 May 2010, University of Victoria, Canada. Data downloaded on 17 June 2013."

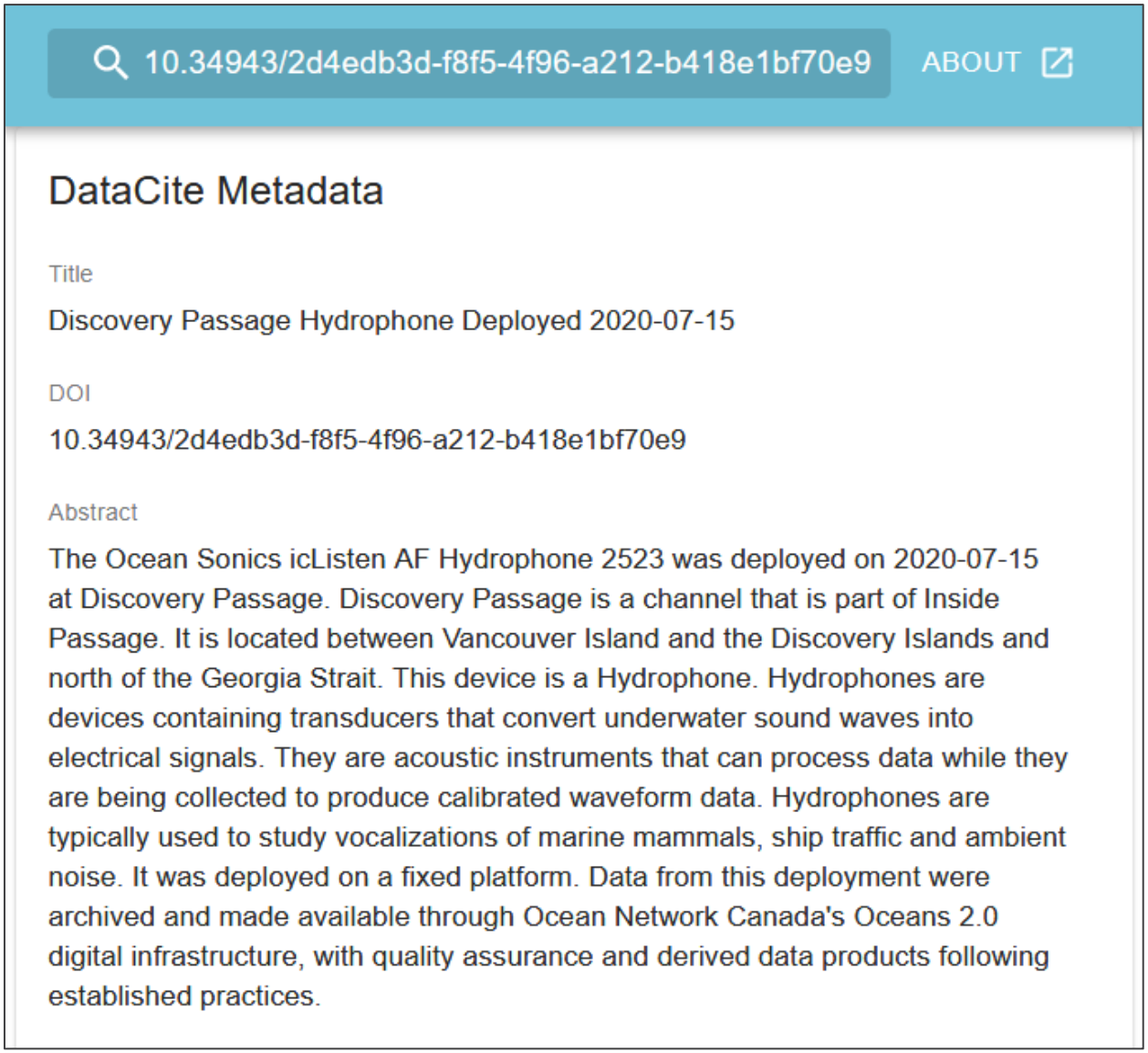
NOW

https://doi.org/10.34943/a53cf191-6916-4d8d-8c6e-f56244016a40

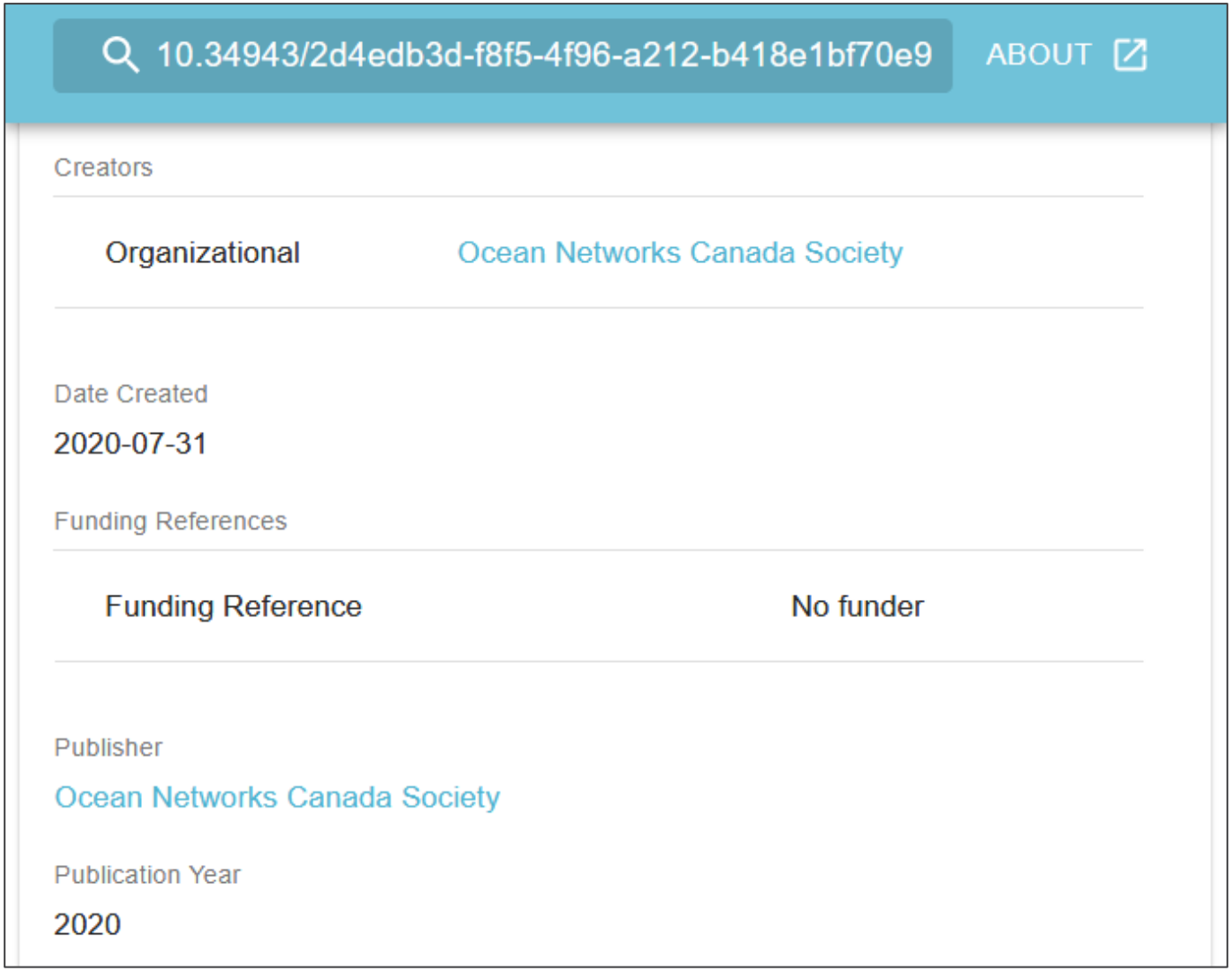
ONC Landing Page - Fixed Position Device



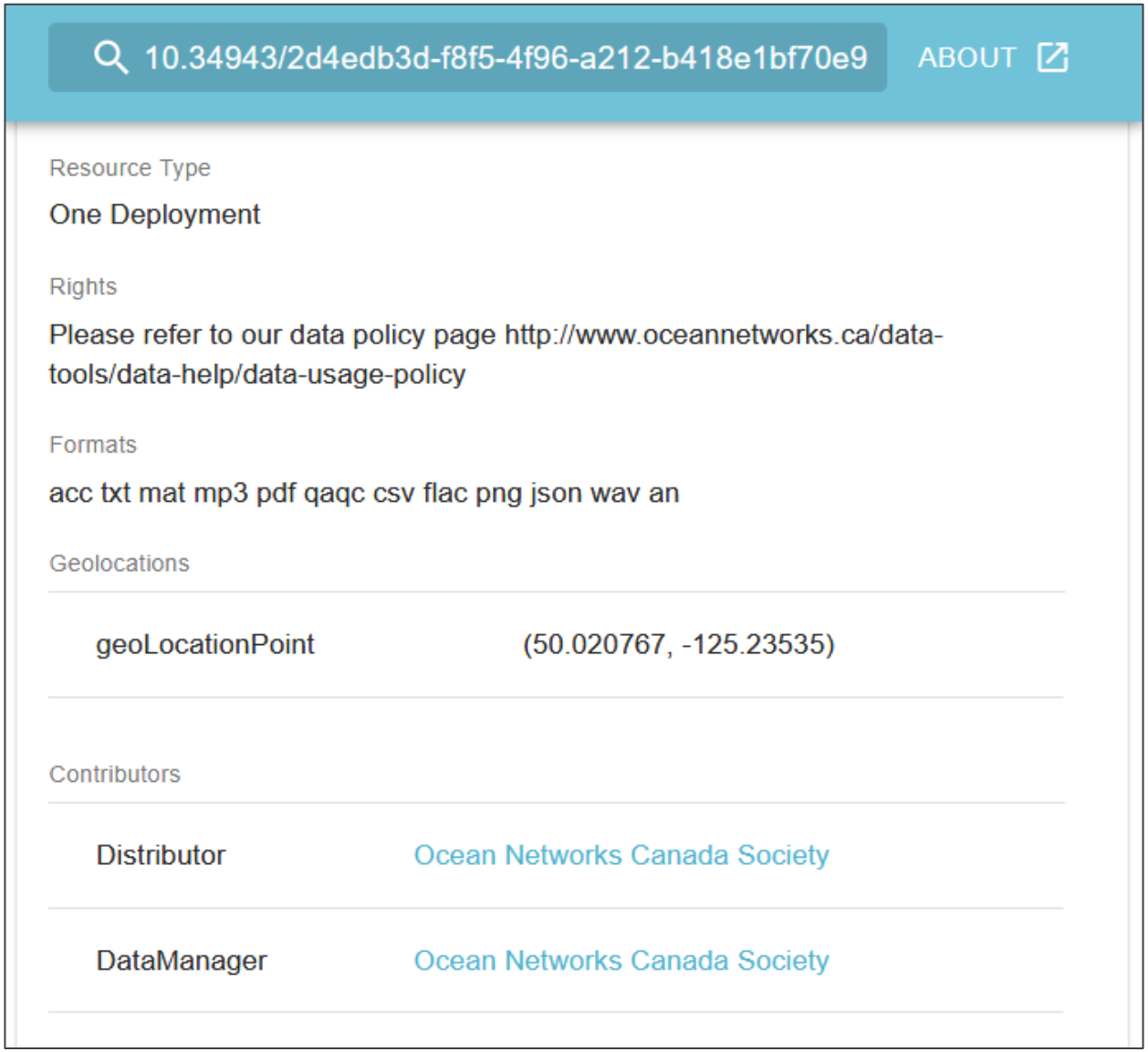
ONC Landing Page - Fixed Position Device



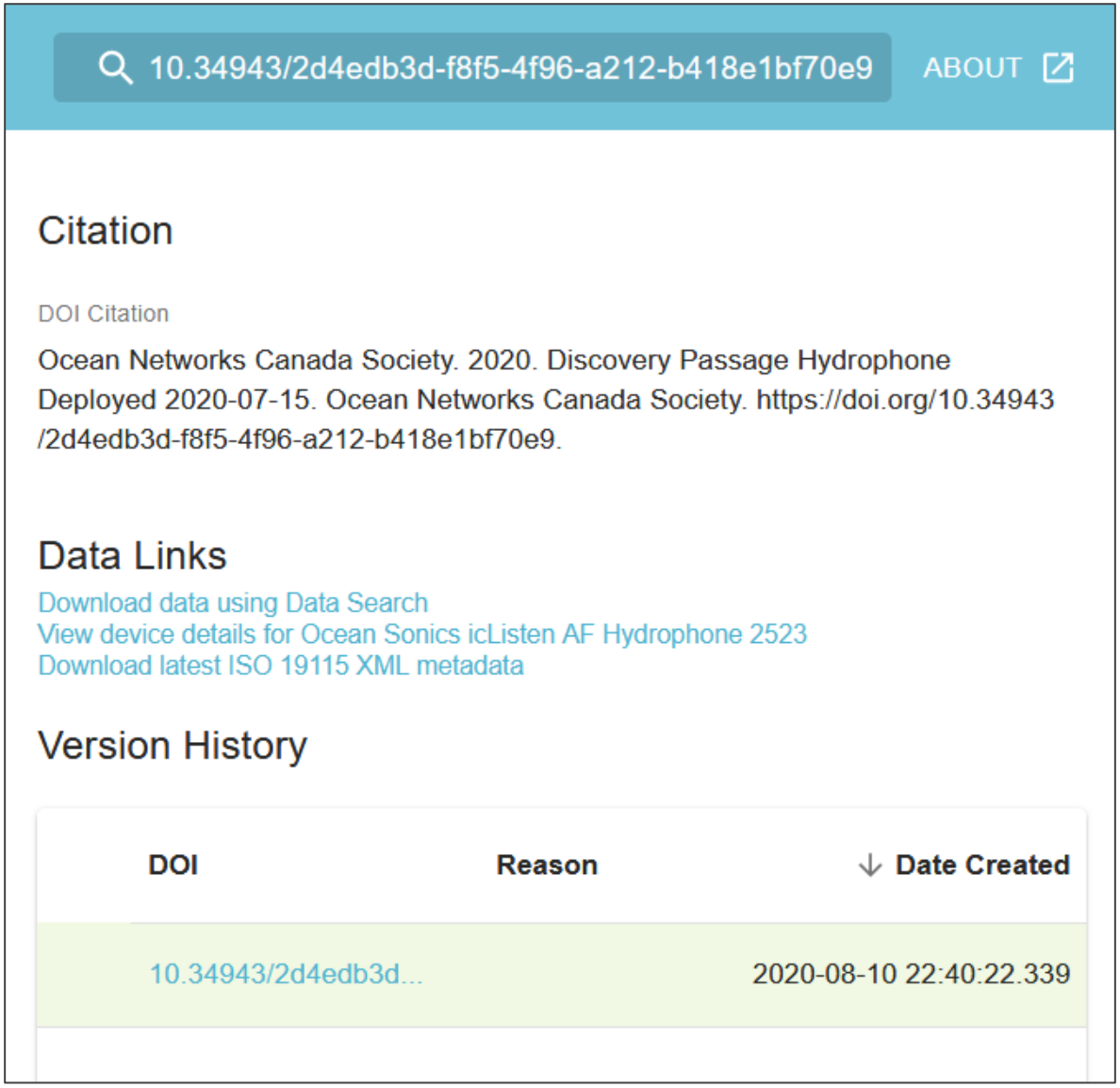
ONC Landing Page - Fixed Position Device



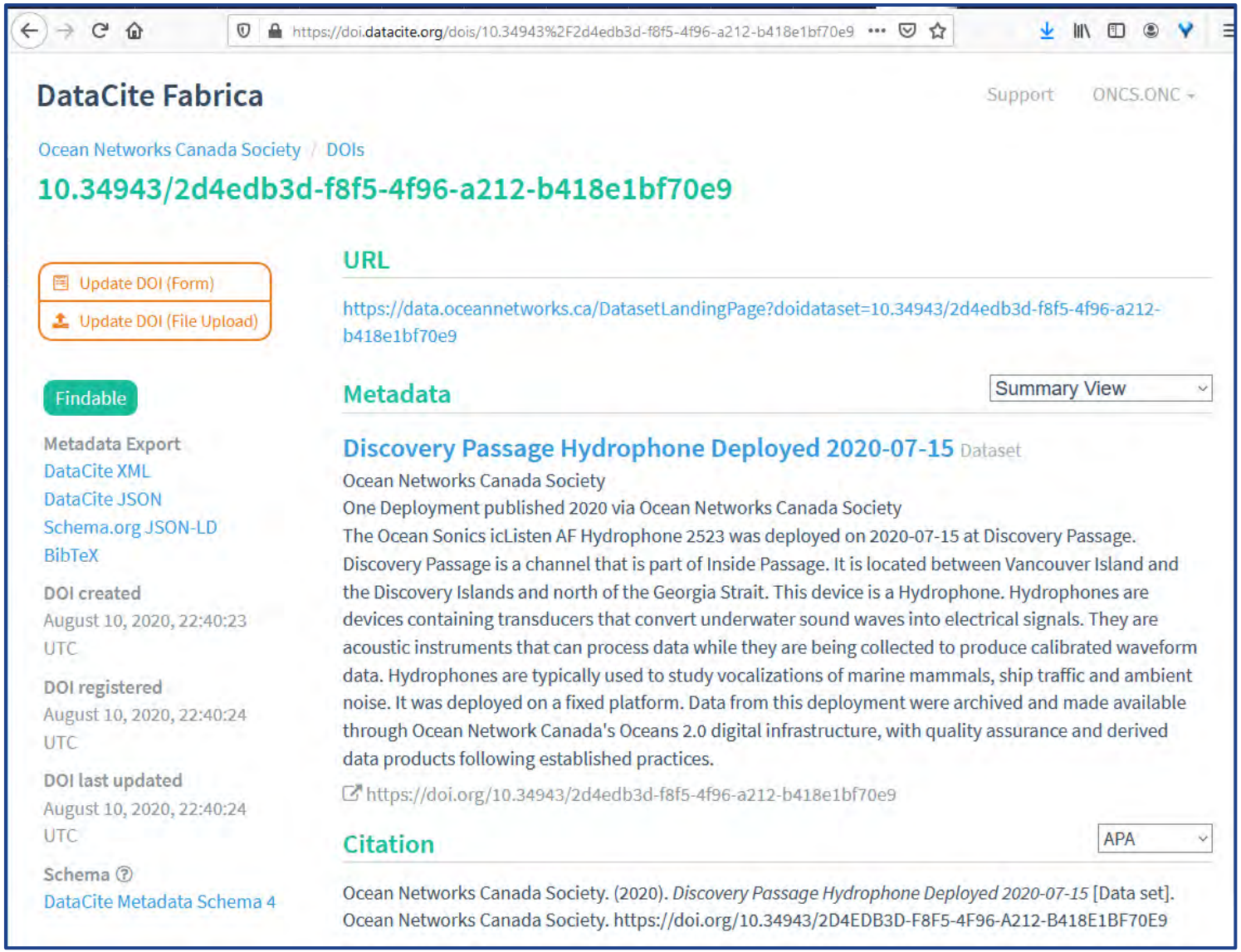
ONC Landing Page - Fixed Position Device



ONC Landing Page - Fixed Position Device



DataCite Metadata Record



Challenges: diverse cases

Supporting the DataCite metadata kernel and ISO 19115 requirements, especially for automatically generated records, took some tweaking of ONC's existing metadata practices - what we stored, where, and how.

- Geospatial metadata: fixed, mobile, and remote sensing instruments
- Data Attributions: data agreement partners with shared responsibility for instruments and their data products
    - including Party Identifiers: ORCID and ROR
- Data Subsets: how to ensure reproducibility by supporting re-creation of previously downloaded subsets of datasets
- API integration: DOIs and Query PIDs, citations, full metadata records
- Versioning: how to display the provenance of evolving datasets

Geospatial case: Mobile Device - *Ferry Thermosalinograph*

- Geographic extent metadata needed for DataCite and ISO 19115
- For mobile devices, **geolocations** field defines a "**bounding box"** instead of a fixed point

10.34943/970f127b-13ba-4b5c-9ac5-f5e9797e711e ABOUT

Formats

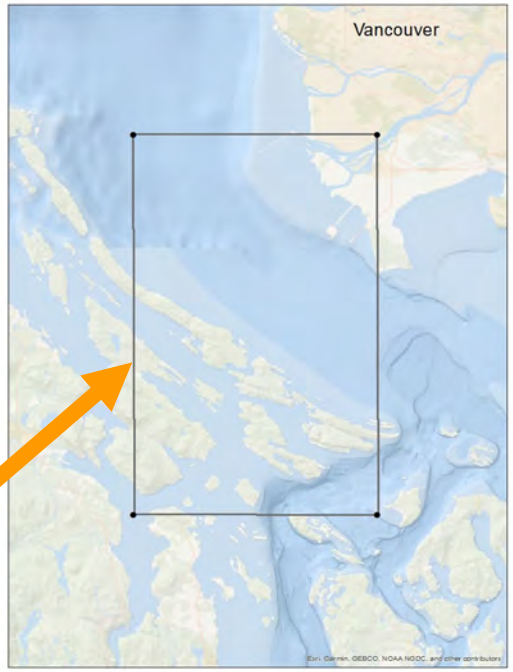
mat txt pdf qaqc csv png json

Geolocations

North Bounding Latitude	49.13045
South Bounding Latitude	48.690001
East Bounding Longitude	-123.082609
West Bounding Longitude	-123.511584

Contributors

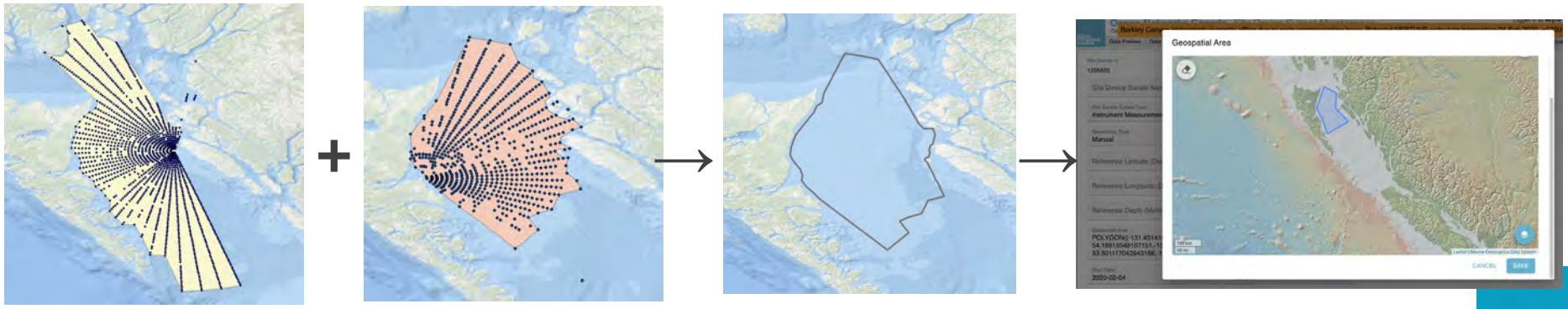
Distributor	Ocean Networks Canada Society
DataManager	Ocean Networks Canada Society



Geospatial case: Remote Sensing Device - *Oceanographic RADAR*

Geographic Extent Metadata

- needed for DataCite and ISO 19115 metadata fields
- existing metadata at ONC only captured **physical location** of instrument, not necessarily representing the **geographic coverage** of dataset
- implemented geographic extent metadata framework (data model, user interface, service integration) for **remote sensing instruments**
- developing procedures for calculating and populating geographic extent for these instruments (e.g., radars, acoustic instruments) – initially manual, but goal is to automate the algorithms based on data acquisition configuration parameters



Responsible Parties - RORs in DataCite

- Support for ROR introduced in *DataCite Metadata Kernel 4.3* (August 2019)
- usable as nameIdentifier or affiliationIdentifier
- new Affiliation subproperties (formerly free text) for Creator and Contributor:
  - affiliationIdentifier
  - affiliationIdentifierScheme
  - schemeURI
- FundingReference property:
  - ROR added as option for funderIdentifier subproperty 'schemeURI'
- Publisher still doesn't support affiliation...yet

Table 3: Expanded DataCite Mandatory Properties

ID	DataCite-Property	Occ	Definition	Allowed values, examples, other constraints
2	Creator	1-n	The main researchers involved in producing the data, or the authors of the publication, in priority order. To supply multiple creators, repeat this property.	May be a corporate/institutional or personal name. Note: DataCite infrastructure supports up to 8000-10000 names. For name lists above that size, consider attribution via linking to the related metadata.
2.4	nameIdentifier	0-n	Uniquely identifies an individual or legal entity, according to various schemas.	The format is dependent upon schema.
2.4.a	nameIdentifierScheme	1	The name of the name identifier schema.	If nameIdentifier is used, nameIdentifierScheme is mandatory. Examples: ORCID[13], ISNI[14], ROR[15], GRID[16].
2.4.b	schemeURI	0-1	The URI of the name identifier schema.	Examples: http://www.isni.org/ https://orcid.org https://ror.org/ https://www.grid.ac/
2.5	affiliation	0-n	The organizational or institutional affiliation of the creator.	Free text. The creator's nameType may be *Organizational* or *Personal*. In case of an organizational creator, e.g. a research group, you can add here the name of the formal institution to which the creator belongs.
2.5.a	affiliationIdentifier	0-1	Uniquely identifies the organizational affiliation of the creator.	The format is dependent upon schema. Examples : https://ror.org/04aj4c181 grid.461819.3
2.5.b	affiliationIdentifierScheme	1	The name of the affiliation identifier schema.	If affiliationIdentifier is used, affiliationIdentifierScheme is mandatory. Examples : ROR, GRID
2.5.c	SchemeURI	1	The URI of the affiliation identifier schema	Examples : http://www.isni.org http:/orcid.org https://ror.org/ https://www.grid.ac/

Data Partner Attributions: RORs in action

- Integrated into Oceans 2.0 data model & UI, DataCite metadata, & Landing Page

Ocean Networks Canada Dataset Landing Page

Oceans 2.0

Data Preview | Data Search | Plotting Utility | SeaTube | Digital Fishers | Cameras | More | Admin

10.34943/5218ac5d-028c-47b8-a381-59979cdc1954 ABOUT

DataCite Metadata

Title

Gascoyne Inlet Conductivity Temperature Depth Deployed 2019-08-19

DOI

10.34943/5218ac5d-028c-47b8-a381-59979cdc1954

Abstract

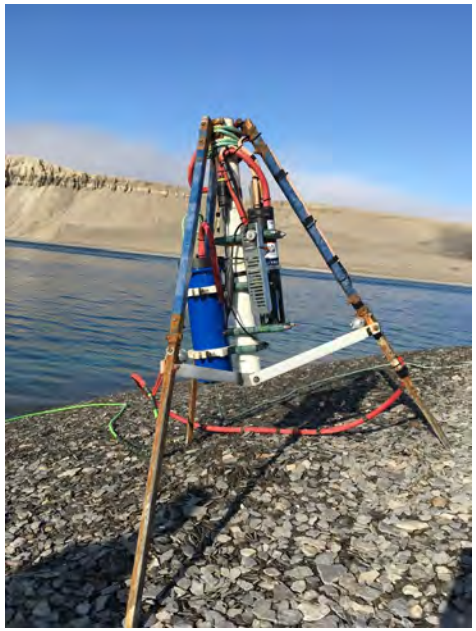
The AML CTD Metrec X 50233 was deployed on 2019-08-19 at Gascoyne Inlet. Gascoyne Inlet is an inlet within Nunavut, on the southwestern tip of Devon Island, and is nearby to Cape Ricketts, Caswall Tower and Cape Liddon. This device is a Conductivity Temperature Depth. Conductivity Temperature Depth (CTD) is an instrument package that contains sensors for measuring the conductivity, temperature, and pressure of seawater. Salinity, sound velocity, depth and density are variables that can be derived from sensor measurements. CTDs can carry additional instruments and sensors such as oxygen sensors, turbidity sensors and fluorometers. It was deployed platform. Data from this deployment were archived and made available through Ocean Network Canad digital infrastructure, with quality assurance and derived data products following established practices.

Creators

Organizational	Defence Research and Development Canada (DRDC)
Organizational	Ocean Networks Canada Society

Date Created

2020-03-10



ROR Search Registry... ABOUT SCOPE

https://ror.org/00hgy8d33

Defence Research and Development Canada

DRDC, RDDC , RECHERCHE & DÉVELOPPEMENT POUR LA DÉFENSE CANADA

WEBSITE
http://www.drdc-rddc.gc.ca/en/index.page

OTHER IDENTIFIERS
GRID grid.1463.0
ISNI 0000000106926582
Crossref Funder ID 501100002956
Wikidata Q1182587

CANADA GOVERNMENT

Data Subsets

What if you don't need the entire dataset?

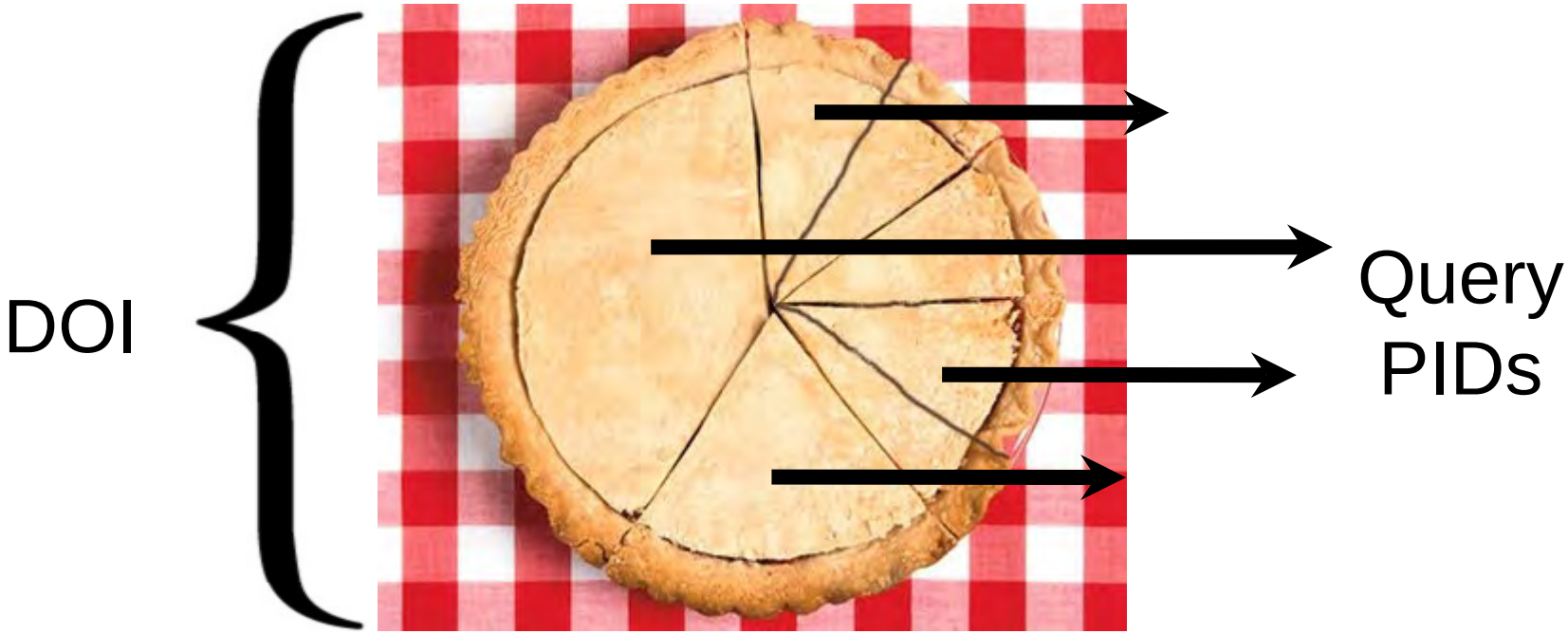
Many subsets of data that can be extracted from a single dataset, defined by: sensor, time range, data product, file format, etc...

In these cases, **citing the DOI of the full dataset is inadequate for reproducibility.**

Recall the 14 recommendations of the *RDA Dynamic Data Citations WG Guidelines* **Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use** (2016):

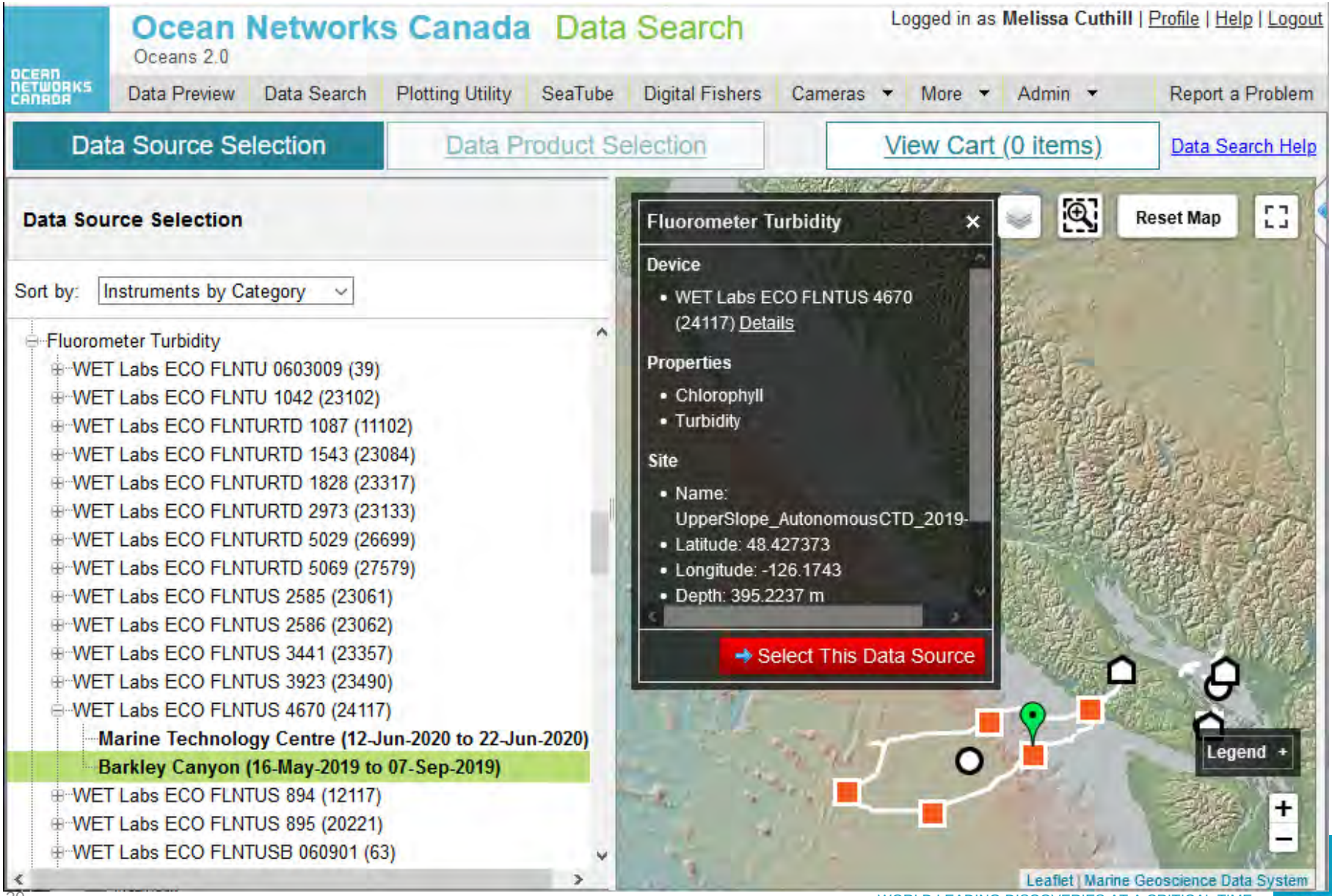
- **R3 - Query Store Facilities**: Provide means for storing queries and the associated metadata in order to re-execute them in the future
- **R7 - Query Timestamping**: Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time)
- **R8 - Query PID**: Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID of the earlier query to the user.
- **R9 - Store the Query**: Store query and metadata (e.g. PID,original and normalised query, query and result set check-sum, timestamp, superset PID, data set description, and other) in the query store.

Enter the Query PID!

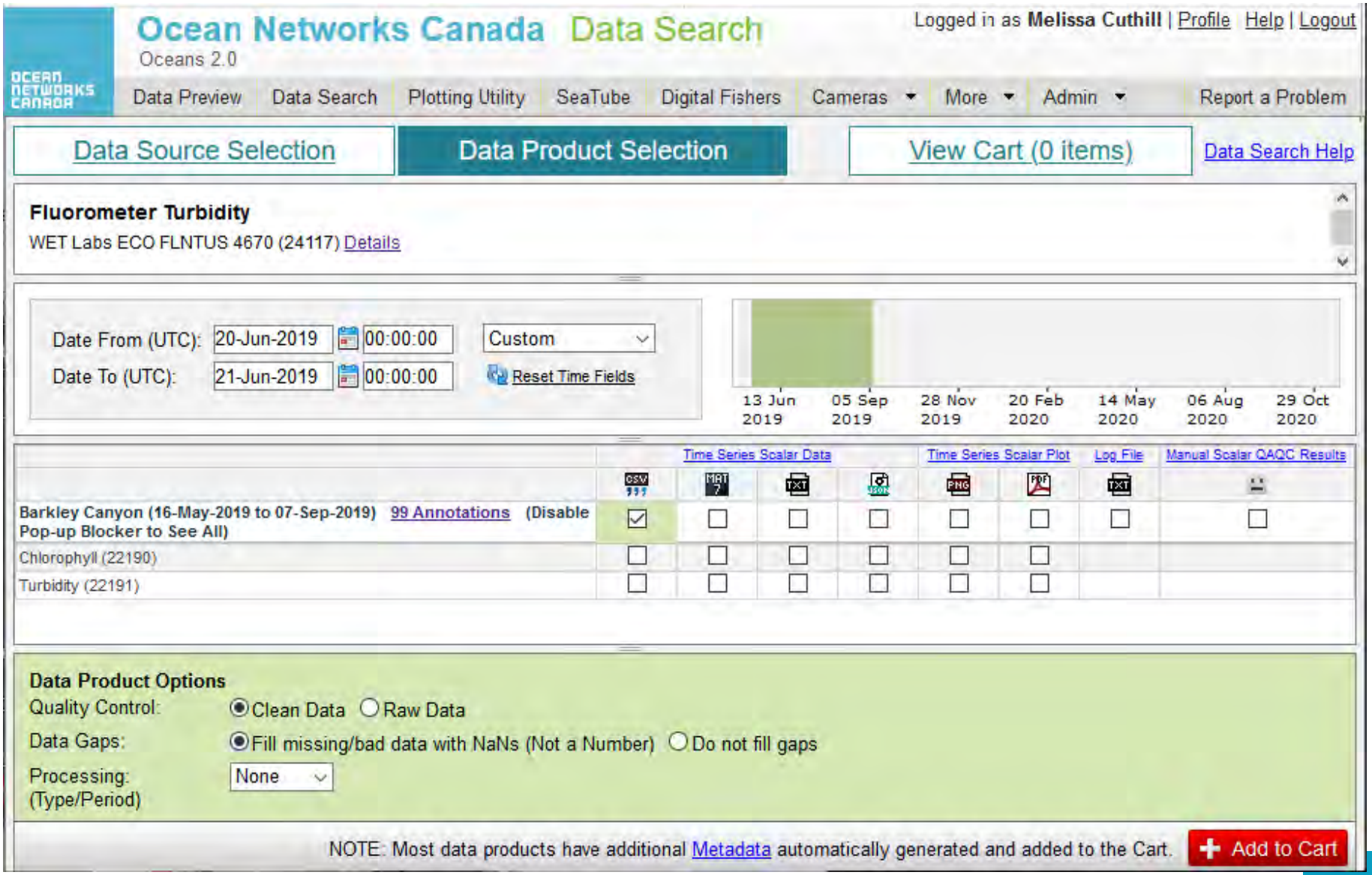


- Every data search (query) in Oceans 2.0 is **saved** in the database and labeled with its own **local identifier**, the '**Query PID**'
- Query PIDs can be used like a DOI in the Oceans 2.0 Landing Page Resolver to view a landing page with **additional details** specific to that search

Data Search in Oceans 2.0: *Dataset selection*



Data Search in Oceans 2.0: *Subset Query Details*



Query PID Landing Page: *Subset Query Details*

Ocean Networks Canada Dataset Landing Page

Oceans 2.0

Logged in as **Melissa Cuthill** | Profile | Help | Logout

Data Preview | Data Search | Plotting Utility | SeaTube | Digital Fishers | Cameras | More | Admin | Request Support | Report a Problem

8510415 ABOUT

DataCite Metadata

Title

Barkley Canyon Upper Slope Fluorometer Turbidity Deployed 2019-05-16

DOI

10.34943/fa04d675-3df2-4dc3-810b-cb365f7ec492

Abstract

The WET Labs ECO FLNTUS 4670 was deployed on 2019-05-16 at Barkley Canyon Upper Slope. Upper Slope is a location within Barkley Canyon, which is located on the upper continental slope. This device is a Fluorometer Turbidity. Fluorometer Turbidity instruments measure chlorophyll fluorescence and turbidity within the same volume of seawater. The instrument uses a light emitting diode (LED) to provide an excitation source. The fluoresced light is received by a detector at a particular angle from the LED source, and uses an interference filter to discriminate against scattered excitation light. Turbidity is measured at the same time, by detecting scattered light from a LED, which is positioned at the same angle as the chlorophyll fluorescence. It was deployed on a fixed platform. Data from this deployment were archived and made available through Ocean Network Canada's Oceans 2.0 digital infrastructure, with quality assurance and derived data products following established practices.

Creators

Organizational	Ocean Networks Canada Society

Date Created

2019-12-16

Funding References

Funding Reference	No funder

Publisher

Ocean Networks Canada Society

Publication Year

Query Details

Data Product

Time Series Scalar Data

Query Date Created

2020-05-08T17:26:27.733Z

Query Date From

2019-06-20T00:00:00.000Z

Query Date To

2019-06-21T00:00:00.000Z

Variables

All

Format

csv

Data Product Options

Data Gaps:	Fill missing/bad data with NaNs (Not a Number)
Quality Control:	Clean Data
Processing: (Type/Period)	Average / 1 Minute

Citation

Query Citation

Ocean Networks Canada Society. 2019. Barkley Canyon Upper Slope Fluorometer Turbidity Deployed 2019-05-16. Ocean Networks Canada Society. https://doi.org/10.34943/fa04d675-3df2-4dc3-810b-cb365f7ec492. Subset Query: 8510415. Accessed 2020-05-08.

Citation Text and Metadata Web Services (API)

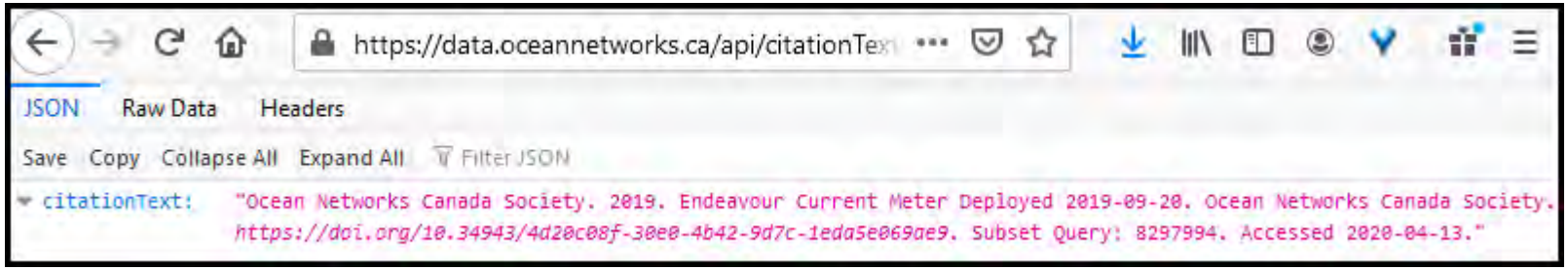
DOIs and Query PIDs can also be used with the **Oceans 2.0 API** to retrieve the data citation formatted according to the **ESIP Data Citation Guidelines for Earth Science Data, v.2**:

DOI

https://data.oceannetworks.ca/api/citationText?method=get&doi=10.34943/115343a9-6d88-4f6c-a88b-9a7b17ad53e0

Query PID

https://data.oceannetworks.ca/api/citationText?method=get&queryPid=8297994



Versioning Data

Each batch is given a name & reason.

Batches contain:

1. Metadata Triggers
2. Data Versioning Tasks
3. DataCite DOI Updates

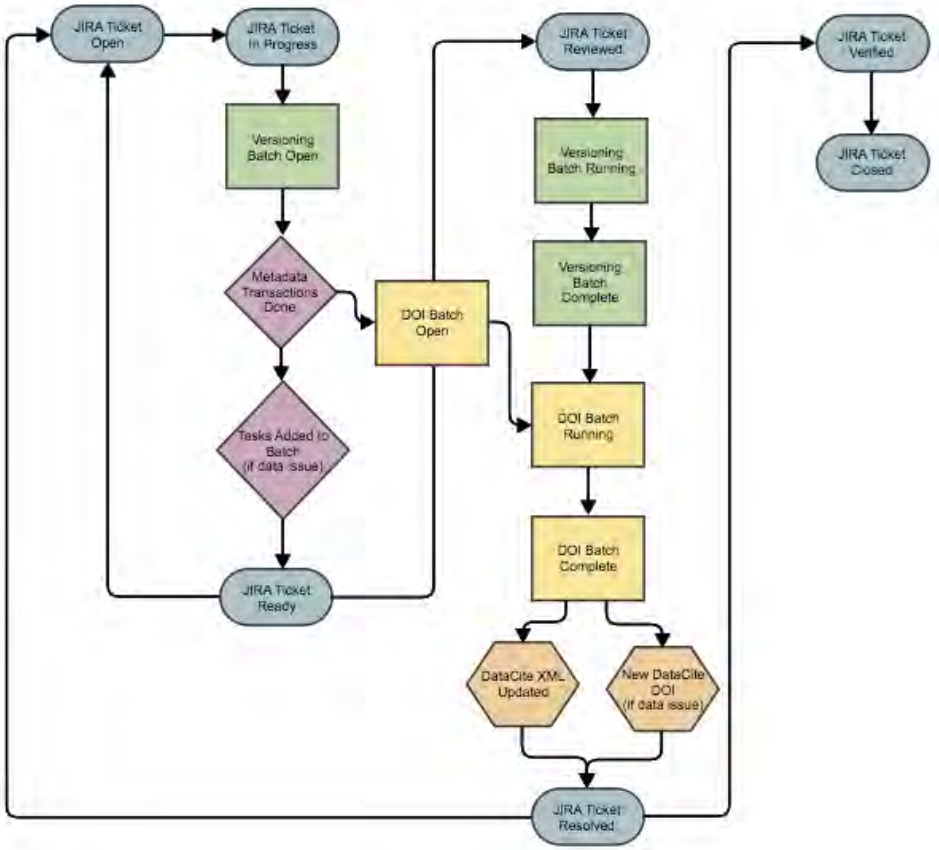
Example metadata triggers:

- calibration formula change
- device attribute change
- parser change

Triggers automatically identify affected datasets, or data steward can manually add datasets affected.

Versioning tasks currently supported are

- reprocessing to parse data (e.g., after formula or parser fix)
- re-postprocessing of derived data products (e..g, after algorithm fix or parameter change)
- file uploads (to fill gaps or replace faulty files)



Versioning Data

DataCite DOI updates include

- generation of new DOI that indicates it is the new version of the prior one
- update to the prior DOI xml to indicate it is now the previous version of the new one

The ONC dataset landing page is updated in the Version History section, with the resolved DOI highlighted, and versioning reasons and triggers provided.

```
<relatedIdentifiers>
        <relatedIdentifier relatedIdentifierType="DOI"
relationType="IsPreviousVersionOf">10.21383/5efd1457-
db3f-45e0-9802-9e7e58edf004</relatedIdentifier>

    </relatedIdentifiers>

<relatedIdentifiers>
        <relatedIdentifier relatedIdentifierType="DOI"
relationType="IsNewVersionOf">10.21383/259fd2ac-
e02d-46a0-a27b-b244b1f46dcb</relatedIdentifier>

    </relatedIdentifiers>
```

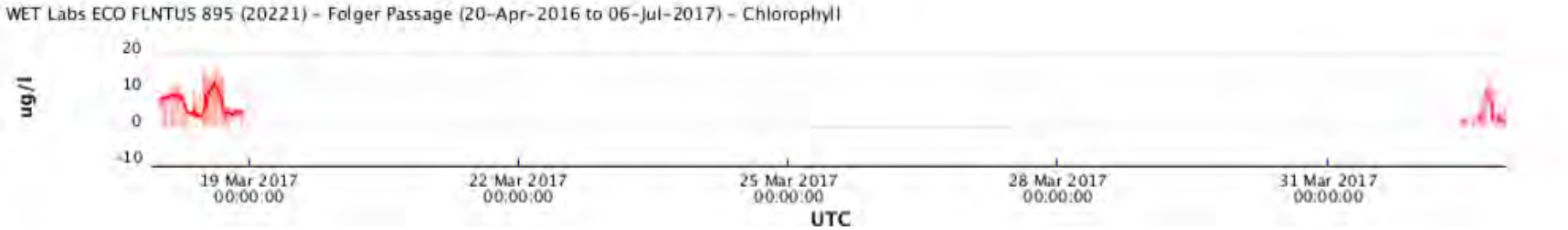
https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.21383/5efd1457-db3f-45e0-9802-9e7e58edf004

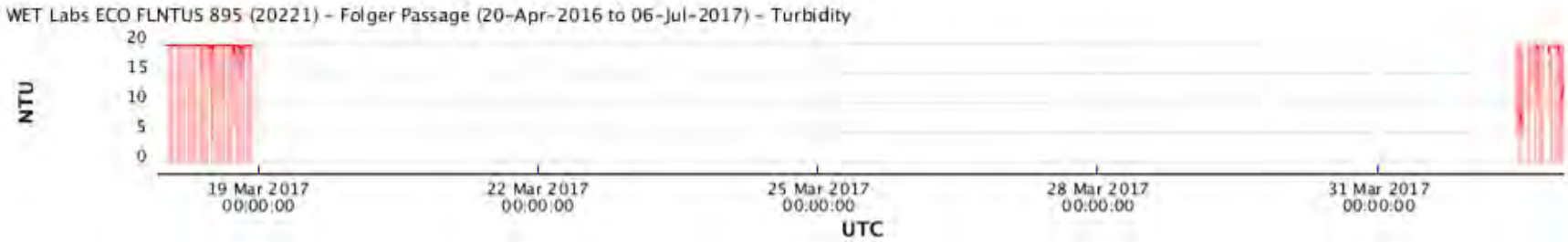
Versioning Data: *Reprocessing*

Use Case: Parsing for sensors on a device were temporarily disabled since the automated QAQC flag system was having issues. This issue was addressed, and then the data have the data reprocessed to fill the gap in our database (raw data was still accumulating, but not being parsed).

Example: WetLabs Fluorometer Turbidity instrument at Folger Pinnacle, deployed 2016-04

Original DOI: https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.34943/bc5bf185-4b4a-4533-963f-2d98dead60ad

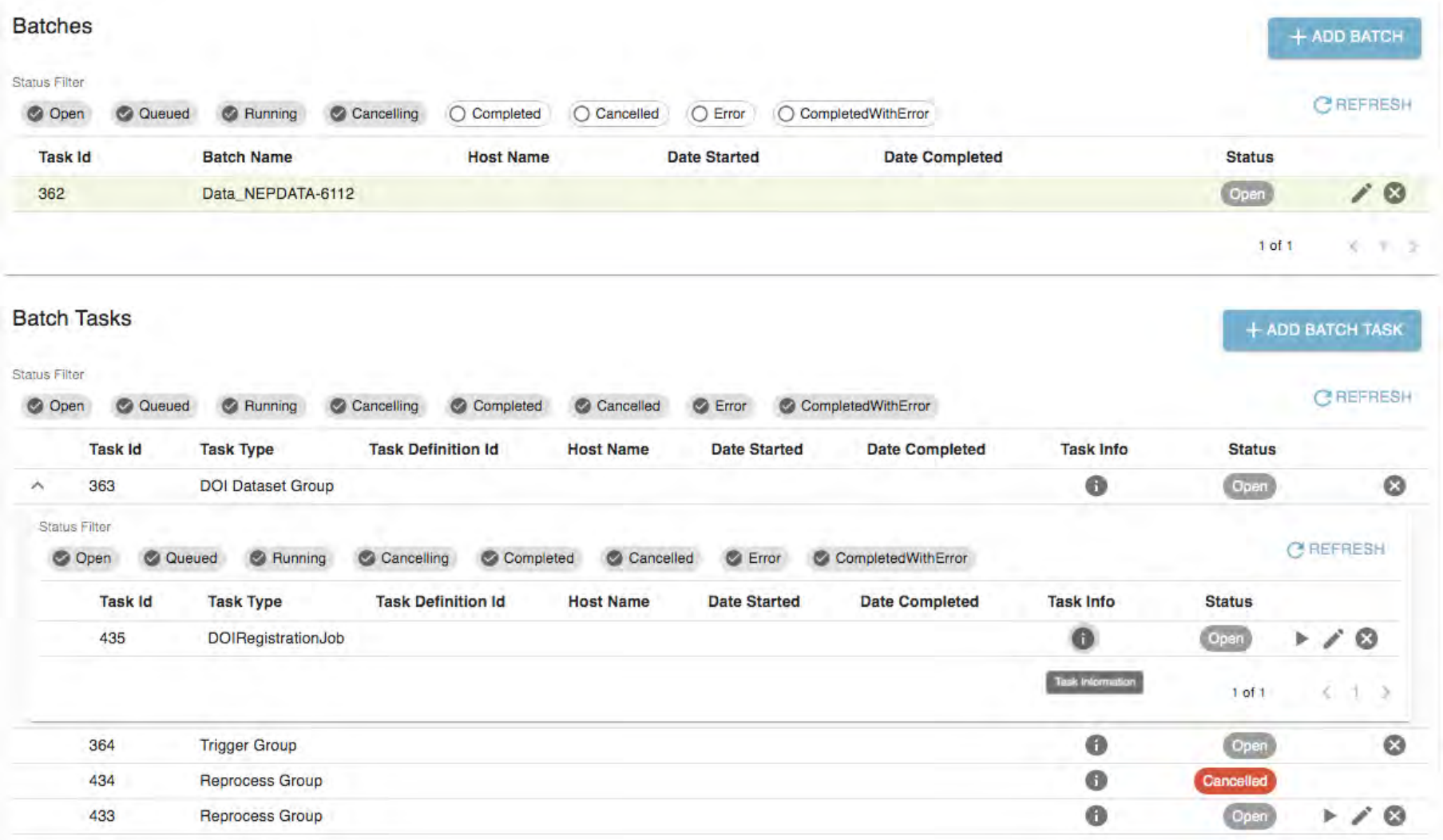




JIRA Ticket created for follow-up action, assigned to Data Steward

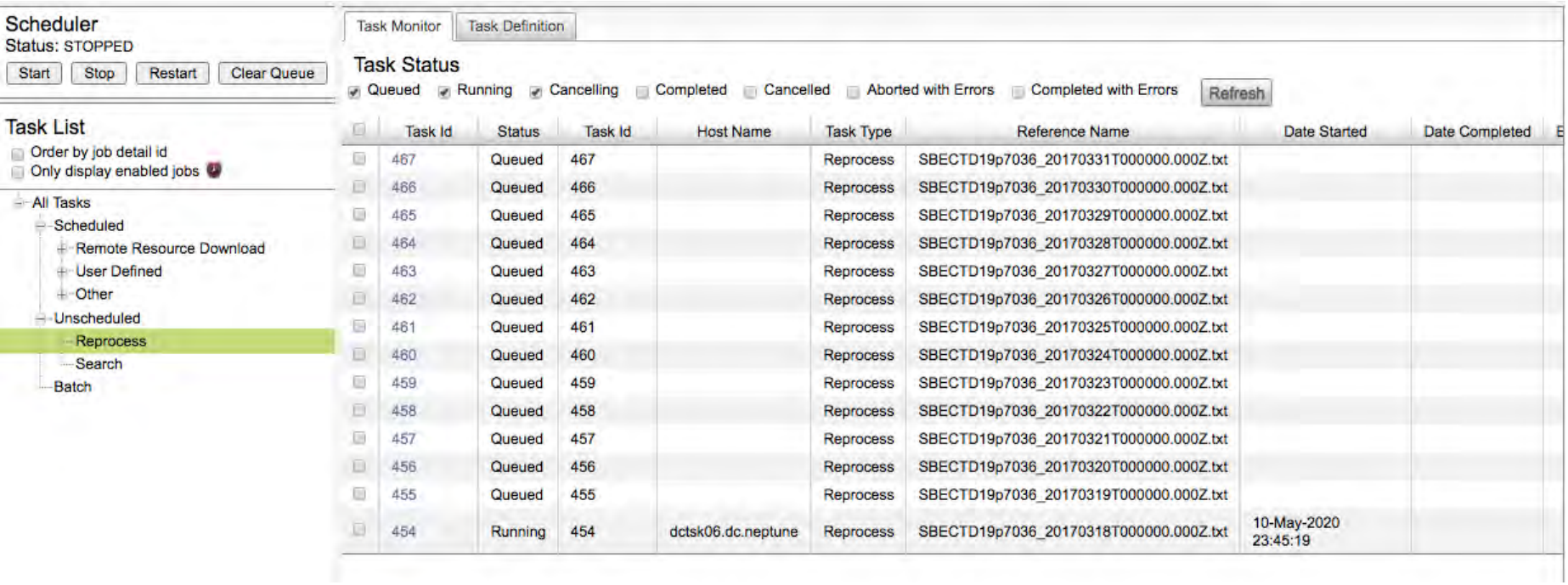
Versioning Data: *Reprocessing*

Batch versioning interface includes support for associating **metadata triggers**, **data versioning** tasks (e.g., reprocessing), and **DataCite DOI updates**.



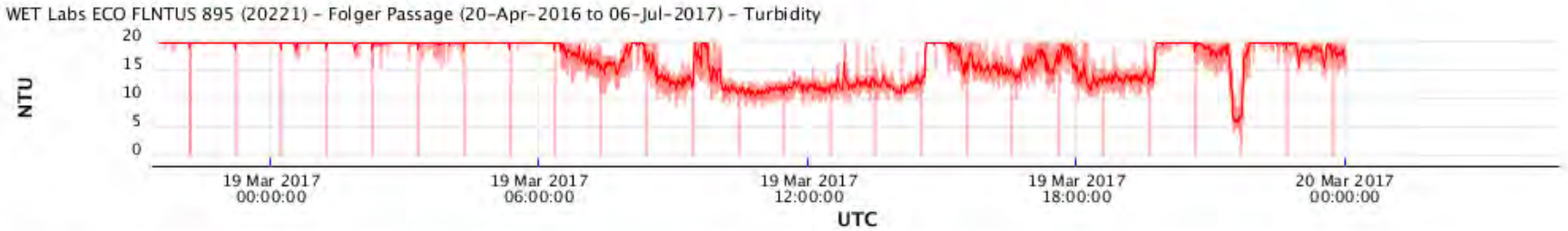
Versioning Data: *Reprocessing*

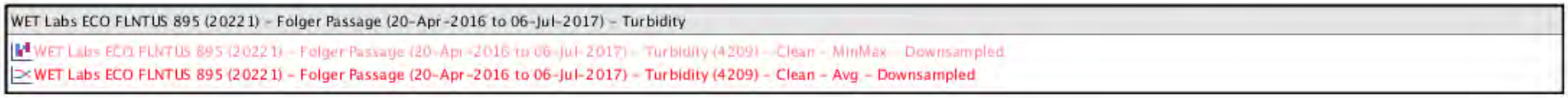
Reprocessing can be monitored as it progresses.



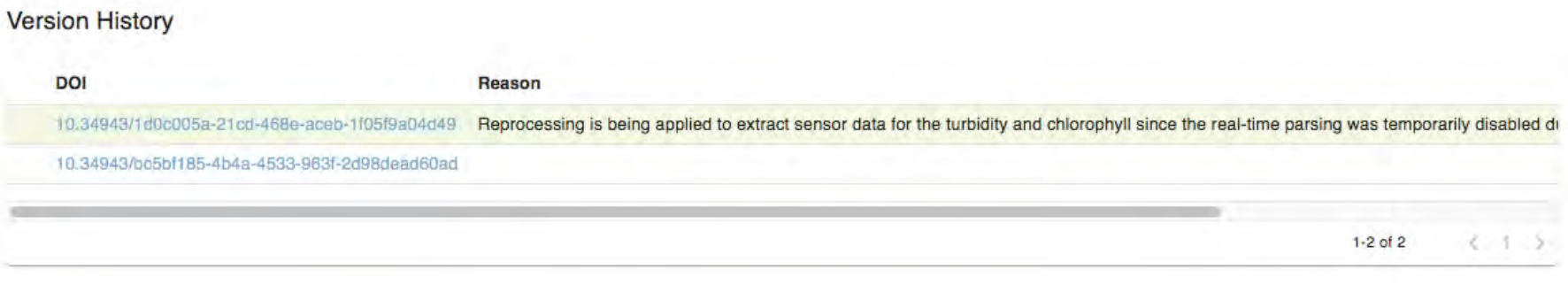
Versioning Data: *Reprocessing*

Gap becoming filled as reprocessing progresses...

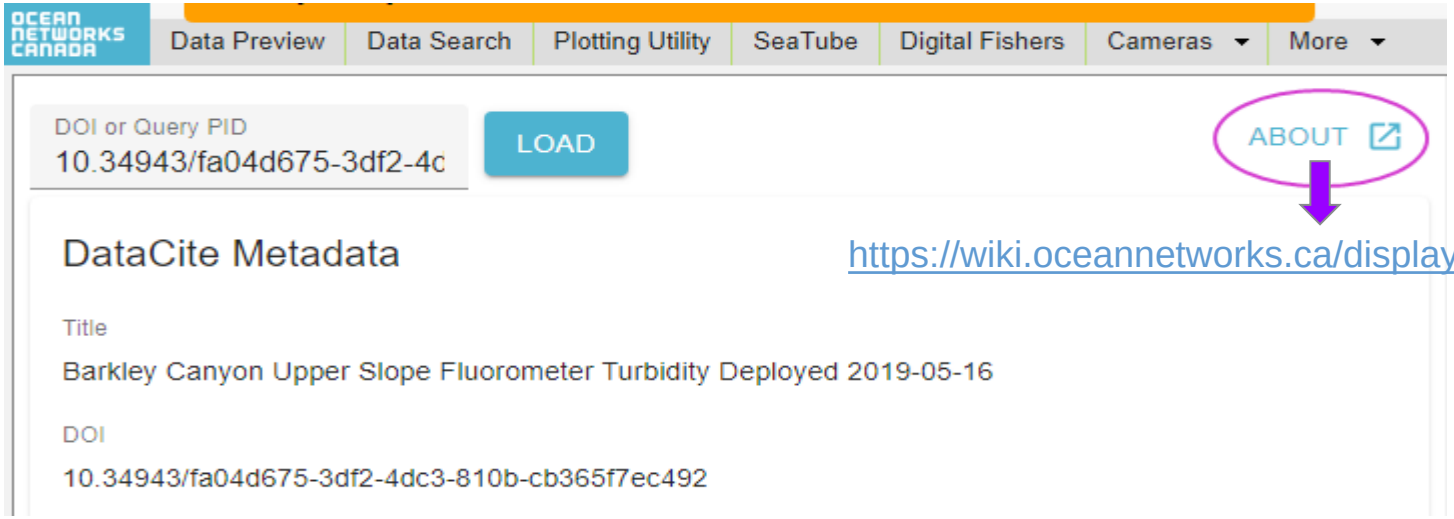




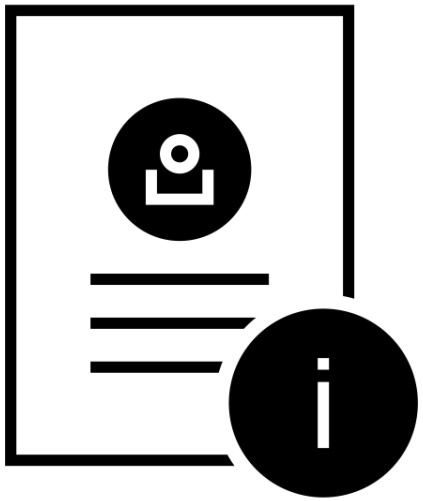
New dataset landing page with provenance information:
https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.34943/1d0c005a-21cd-468e-aceb-1f05f9a04d49



User Documentation



https://wiki.oceannetworks.ca/display/DP/Data+Citations



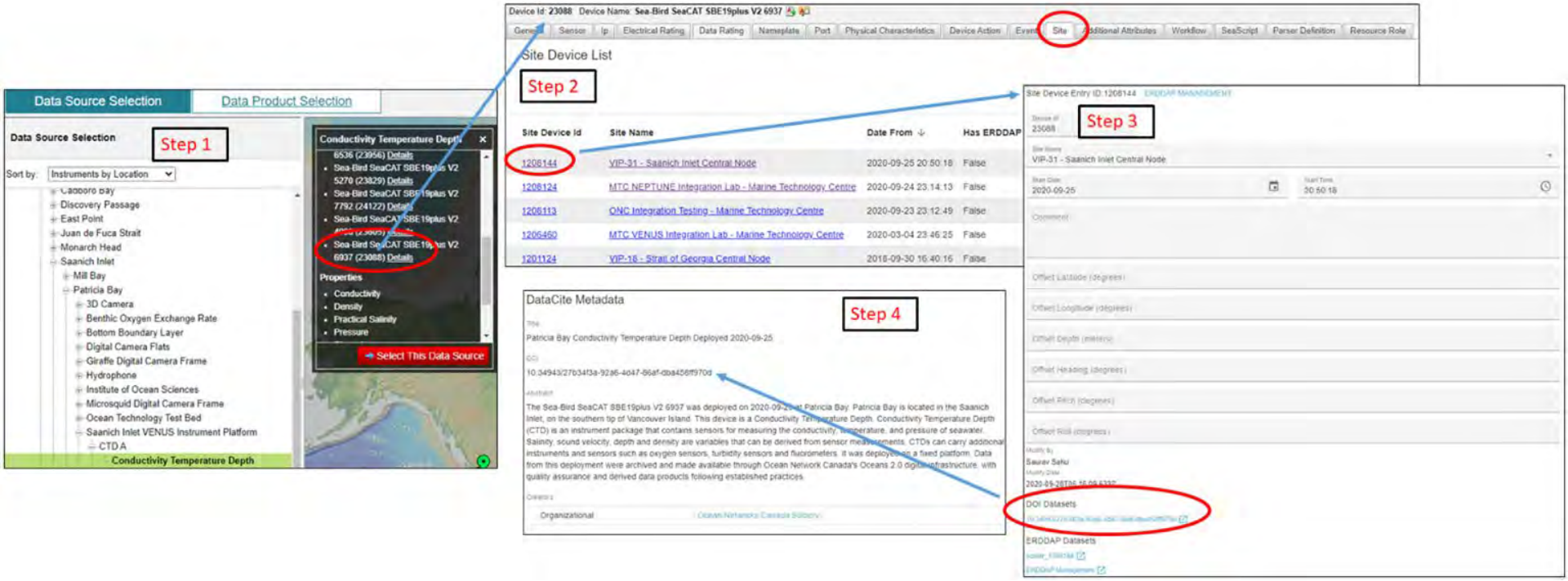
Created by P Thanga Vignesh
from Noun Project

Data Citations

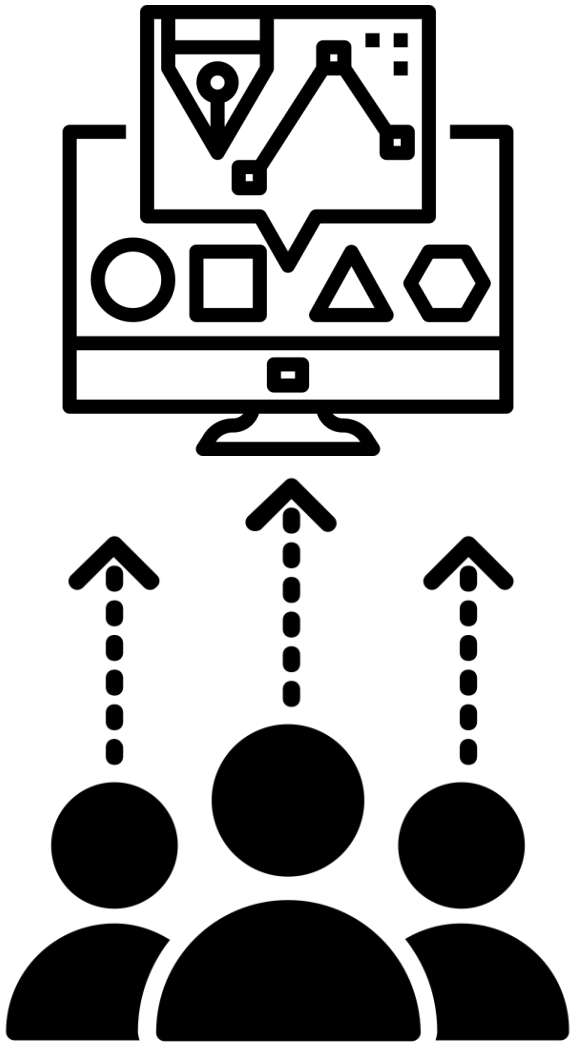
Created by Reyna Jenkyns , last modified by Chantel M Ridsdale on 06-May-20

- What is Data Citation and Why Is It Important?
  - Dynamic Data Citation
- MINTED Project Overview
  - Benefits of MINTED
  - Landing Page
    - → Landing Page Metadata Sections:
  - Query PID
    - → Query-PID Specific Metadata Sections (green box):
- Web Services
  - → Data Citation Text Service
  - → Dataset Metadata Service
- Future Plans
- Recommended Resources

Landing Page Discovery



Data Contributor Documentation



Created by ProSymbols
from Noun Project

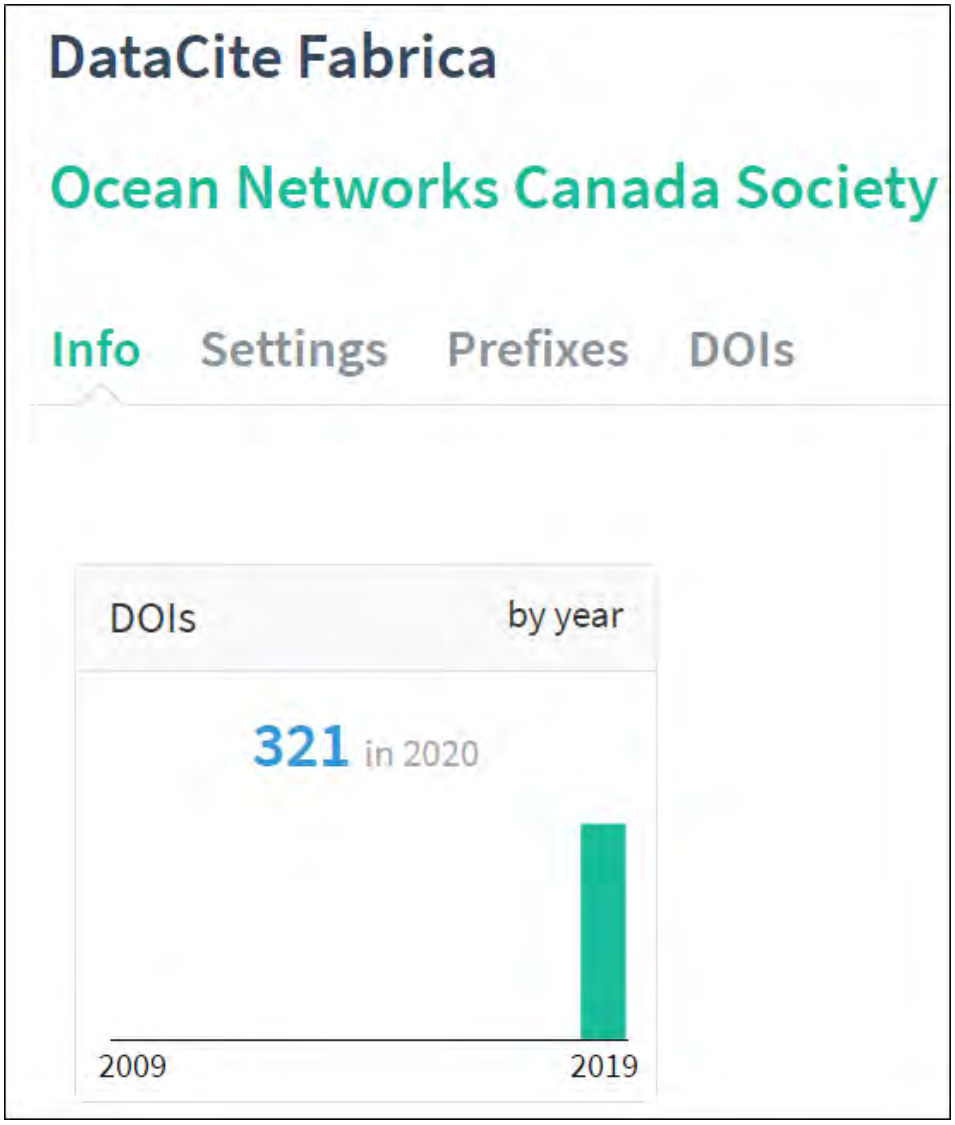
Data Partnerships Home

Created by Reyna Jenkyns , last modified by Chantel M Ridsdale on 18-Aug-20

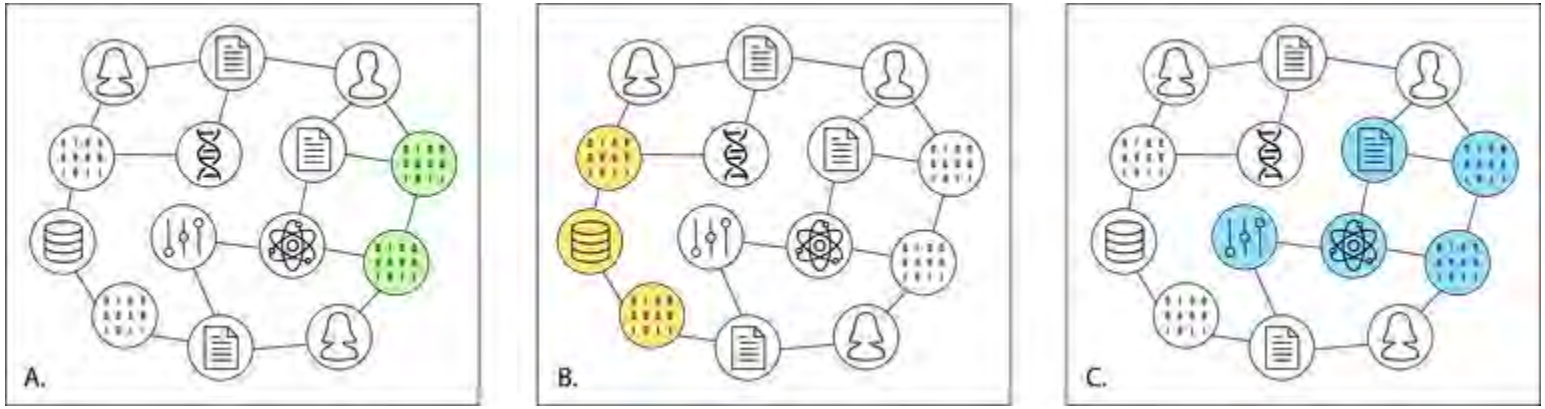
- Overview
- Data Agreements
- Attributions & Credit
  - Data Citations
  - ISO 19115 Metadata Records
    - Individual Metadata Record
    - Metadata Catalog
  - Roles
  - Research Organization Registry
  - Metrics
- Contact Information

https://wiki.oceannetworks.ca/display/DataPartners/Data+Partnerships+Home

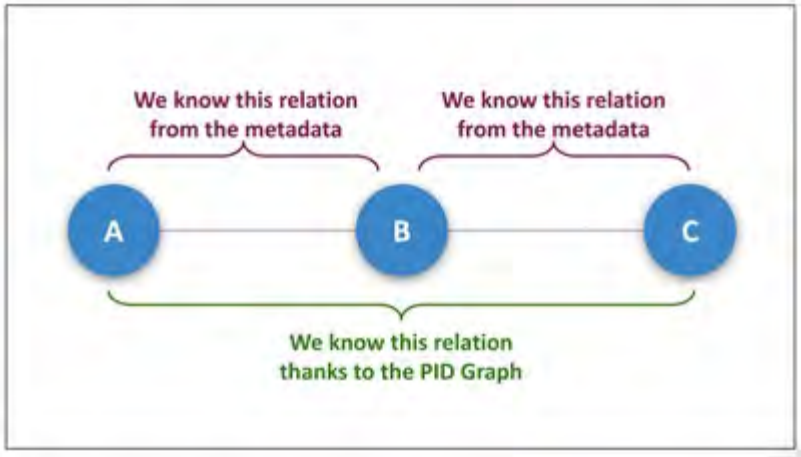
Metrics: DataCite Stats



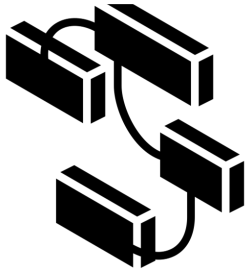
PID Graph

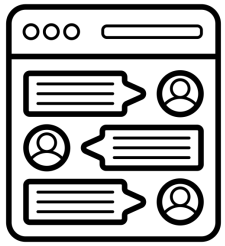


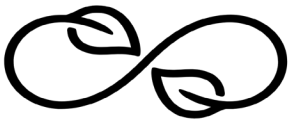
- Courtesy of the FREYA project
- Leverages all the different types of PIDs to unambiguously cross-link research outputs



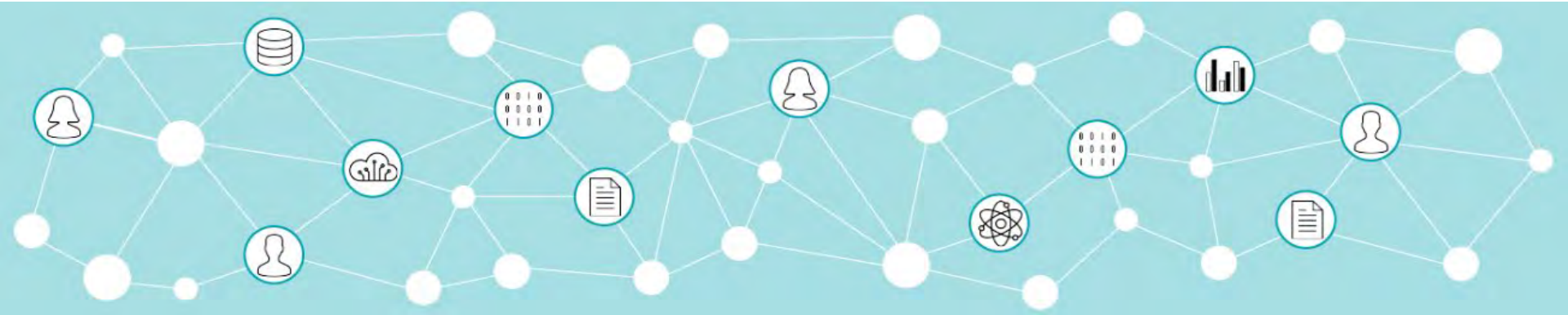
FREYA recognition







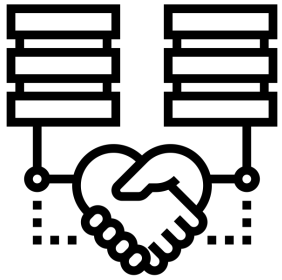
ODIN and THOR



Future Plans

Follow-on project proposal planned to be submitted for additional features in fall 2020 - scope definition in progress...

ISO 19115:2014




```
<relatedIdentifiers>
      <relatedIdentifier relatedIdenifierType=”URL”
relationType=...>
```

OCEAN NETWORKS CANADA

Questions?

reyna@uvic.ca