



**Data Citation
Working Group Mtg @ P20**

March 22 2023, Gothenberg

Andreas Rauber, Mark Parsons

research data sharing without barriers

rd-alliance.org

Agenda

2

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New directions”:
 - Challenges in complex citations
- Other issues, next steps

Welcome!

to the maintenance meeting
of the
WGDC

Agenda

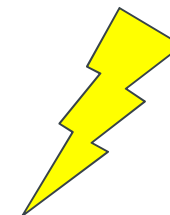
4

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New directions”:
 - Challenges in complex citations
- Other issues, next steps

Challenge: States of Dynamic Data

5

- Usually, datasets have to be static
 - Fixed set of data, no changes:
no corrections to errors, no new data being added
- But: (research) data is **dynamic**
 - Adding new data, correcting errors, enhancing data quality, ...
 - Changes sometimes highly dynamic, at irregular intervals
- Current approaches
 - Identifying entire data stream, without any versioning
 - Using “accessed at” date
 - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to identify precisely the **data as it existed at a specific point in time**



Challenge: Granularity of Subsets

6

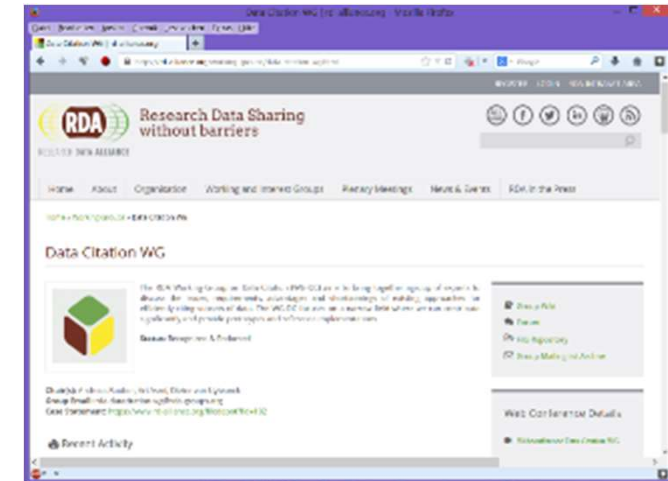
- What about the **granularity** of data to be identified?
 - Enormous amounts of CSV data
 - Researchers use specific subsets of data
 - Need to identify precisely the subset used
 - Current approaches
 - Storing a copy of subset as used in study -> scalability
 - Citing entire dataset, providing textual description of subset -> imprecise (ambiguity)
 - Storing list of record identifiers in subset -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to identify precisely the **subset of (dynamic) data used** in a process



RDA WG Data Citation



- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – September 2015
 - Concentrating on the problems of **large, dynamic (changing) datasets**
- Final version presented Sep 2015 at P7 in Paris, France
- Endorsed September 2016 at P8 in Denver, CO
- Since: support for take-up/adoption, lessons-learned



<https://www.rd-alliance.org/groups/data-citation-wg.html>

Dynamic Data Identification and Citation



We have: Data + Means-of-access (“query”)

Dynamic Data Identification and Citation



We have: Data + Means-of-access (“query”)

**Dynamic Data Citation:
Cite (dynamic) data dynamically via query!**

Dynamic Data Identification and Citation



We have: Data + Means-of-access (“query”)

**Dynamic Data Citation:
Cite (dynamic) data dynamically via query!**

Steps:

1. Data → versioned (history, with time-stamps)

Dynamic Data Identification and Citation



We have: Data + Means-of-access (“query”)

**Dynamic Data Citation:
Cite (dynamic) data dynamically via query!**

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

Dynamic Data Identification and Citation



We have: Data + Means-of-access (“query”)

**Dynamic Data Citation:
Cite (dynamic) data dynamically via query!**

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & assign PID to “QUERY”**, enhanced with

- **Time-stamping** for re-execution against versioned DB
- **Re-writing** for normalization, unique-sort, mapping to history
- **Hashing** result-set: verifying identity/correctness

leading to landing page

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
 - Data (package, access API, ...)
 - PID (e.g. DOI) (Query is time-stamped and stored)
 - Hash value computed over the data for local storage
 - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
 - Provides detailed metadata, link to parent data set, subset,...
 - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

14

- Note: query string provides excellent provenance information on the data set!
- subset of data user gets
 - Data (package, access API, ...)
 - PID (e.g. DOI) (Query is time-stamped and stored)
 - Hash value computed over the data for local storage
 - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
 - Provides detailed metadata, link to parent data set, subset, ...
 - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

15

- Note: query string provides excellent subset of data
- provenance information on the data set! er gets
 - Data (pac
 - PID (e.g.
 - Hash valu
 - Recommended citation text (e.g. PID (EX)
- PID resolves to landing page
 - Provides detailed metadata, link to parent data set, subset,...
 - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

Data Citation – Deployment

16

- Note: query string provides excellent subset of data
- provenance information on the data set! user gets
 - Data (package)
 - PID (e.g. DOI)
 - Hash value
 - Recommended citation text (e.g. PID:RX)
- PID resolves
 - Provides details
 - Option to retrieve
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

Identify which parts of the data are used. If data changes, identify which queries (studies) are affected



Preparing Data & Query Store

- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID

- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification



- **14 Recommendations** grouped into 4 phases:
- **2-page flyer**
<https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html>
- **Detailed report: Bulletin of IEEE TCDL 2016**
http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
- **Adopter's reports, webinars**
<https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html>
- **Review / Lessons Learned**
Andreas Rauber et al., Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data
Harvard Data Science Review, 3(4), 2021.
DOI [10.1162/99608f92.be565013](https://doi.org/10.1162/99608f92.be565013).



HDSR Paper: From Principles to Adoption ¹⁹

Andreas Rauber, Bernhard Gößwein, Carlo Maria Zwölf, Chris Schubert, Florian Wörister, James Duncan, Katharina Flicker, Koji Zettsu, Kristof Meixner, Leslie D. McIntosh, Reyna Jenkyns, Stefan Pröll, Tomasz Miksa, and Mark A. Parsons: **Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data.** Harvard Data Science Review (HDSR), 3(4), 2021.

DOI [10.1162/99608f92.be565013](https://doi.org/10.1162/99608f92.be565013)

- Principles
- 4 Reference implementations
- 8 Adoptions as Case Studies
- Lessons Learned

HDSR Volume 3 Issue 4
DOI: 10.1162/99608f92.be565013
ISSN: 2644-2353

Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data
Andreas Rauber¹, Bernhard Gößwein^{1,3}, Carlo Maria Zwölf⁴, Chris Schubert⁵, Florina Wörister¹, James Duncan⁶, Katharina Flicker¹, Koji Zettsu⁷, Kristof Meixner¹, Leslie D. McIntosh⁸, Reyna Jenkyns⁹, Stefan Pröll¹⁰, Tomasz Miksa^{1,11}, Mark A. Parsons⁷

¹ TU Wien, Vienna, Austria
² University of Alabama in Huntsville, AL, USA
³ Earth Observation Data Centre, Vienna, Austria
⁴ LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne University, UPMC Univ Paris, Meudon, France
⁵ Climate Change Centre Austria, Vienna, Austria
⁶ Forest Ecosystem Monitoring Cooperative, University of Vermont, Burlington, VT, USA
⁷ National Institute of Information and Communications Technology, Tokyo, Japan
⁸ Ripeta, Saint Louis, MO, USA
⁹ Ocean Networks Canada, University of Victoria, Victoria, BC, Canada
¹⁰ Cropster, Innsbruck, Austria
¹¹ SBA Research, Austria

Abstract

Precisely identifying arbitrary subsets of data so that these can be reproduced is a daunting challenge in data-driven science, the more so if the underlying data source is dynamically evolving. Yet, an increasing number of settings exhibit exactly those characteristics: larger amounts of data being continuously ingested from a range of sources (be it sensor values, [online] questionnaires, documents, etc.), with error correction and quality improvement processes adding to the dynamics. Yet, for studies to be reproducible, for decision-making to be transparent, and for meta studies to be performed conveniently, having a precise identification mechanism to reference, retrieve, and work with such data is essential. The Research Data Alliance (RDA) Working Group on Dynamic Data Citation has published 14 recommendations that are centered around time-stamping and versioning evolving data sources and identifying subsets dynamically via persistent identifiers that are assigned to the queries selecting the respective subsets. These principles are generic and work for virtually any kind of data. In the past few years numerous repositories around the globe have implemented these recommendations and deployed solutions. We provide an overview of the recommendations, reference implementations, and pilot systems deployed and then analyze lessons learned from these implementations. This article provides a basis for institutions and data stewards considering adding this functionality to their data systems.

1 Introduction

Accountability and transparency in automated decisions (ACM US Public Policy Council, 2017) have important implications on the way we perform studies, analyze data, and prepare the basis for data-driven decision making. Specifically, reproducibility in various forms, that is, the ability to recompute analyses and arrive at the same conclusions or insights is gaining importance. This has impact on the way analyses are being performed, requiring processes to be documented and code to be shared. More critically, data being the basis of such analyses and thus likely the most relevant ingredient in any data-driven, decision-making process needs to be findable and accessible if any result is to be verified. Yet, identifying precisely which data were used in a specific analysis is a nontrivial challenge in most settings: Rather than relying on static, archived data collected and frozen in time for analysis, today's decision-making processes rely increasingly on continuous data streams that should be available and usable on a continuous basis. Working on last year's (or last week's) data is not an acceptable alternative in many settings. Data undergo complex preprocessing routines, are recalibrated, and data quality is continually improved by correcting error. Thus, data are often in a constant state of flux.

Additionally, data are getting 'big': Enormous volumes of data are being collected, of which specific subsets are selected for analysis, be they a small number of individual values to massive subsets of even bigger data sets. Describing which subset was actually being used and trying to re-create the exact same subset later based on that description may constitute a daunting challenge due to the complexity of subset selection processes (such

1

Large Number of Adoptions

20

- **Standards / Reference Guidelines / Specifications:**
 - Joint Declaration of Data Citation Principles:
Principle 7: Specificity and Verifiability (<https://www.force11.org/datacitation>)
 - ESIP:Data Citation Guidelines for Earth Science Data Vers. 2 (P14)
 - ISO 690, Information and documentation - Guidelines for bibliographic references and citations to information resources (P13)
 - EC ICT TS5 Technical Specification (pending) (P12)
 - DataCite Considerations (P8)
- **Reference Implementations**
 - MySQL/Postgres (P5, P6)
 - CSV files: MySQL, Git (P5, P6, P8, Webinar)
 - XML (P5)
 - CKAN Data Repository (P13)
 - SPARQL (P17)

Large Number of Adoptions

21

- **Early pilot implementations, use cases**
 - DEXHELPP: Social Security Records (P6)
 - NERC: ARGO Global Array (P6)
 - LNEC: River dam monitoring (P5)
 - CLARIN: Linguistic resources, XML (P5)
 - MSD: Million Song Database (P5)
 - many further individual ones discussed ...

Large Number of Adoptions

22

■ Adoptions deployed

- CBMI: Center for Biomedical Informatics, WUSTL (P8, Webinar)
- VMC: Vermont Monitoring Cooperative (P8, Webinar)
- CCCA: Climate Change Center Austria (P10/P11/P12, Webinar)
- EODC: Earth Observation Data Center (P14, Webinar)
- VAMDC: Virtual Atomic and Molecular Data Center (P8/P10/P12, Webinar)
- Ocean Networks Canada (P12, Webinar)
- DBRepo (P19)

■ In progress

- NICT Smart Data Platform (P10/P14)
- Dendro System (P13)
- Deep Carbon Observatory (P12)

Lessons Learned as an FAQ (1 of 2)

23

- **Do the recommendations work for any kind of data?**
Yes, it appears so.
- **Do all updates need to be versioned?**
Ideally, yes. In practice, probably not.
- **May data be deleted?**
Yes, with caution and documentation.
- **What types of queries are permitted?**
Any that a repository can support over time.
- **Does the system need to store every query?**
No, just the relevant queries.
- **Which PID system should be used?**
The one that works best for your situation.
- **When multiple distributed repositories are queried, do we need complex time synchronization protocols?**
No, not if the local repositories maintain timestamps.

Lessons Learned as an FAQ (2 of 2)

24

- **How does this support giving credit and attribution?**
By including a reference to the overall data set as well as the subset.
- **How does this support reproducibility and science?**
By providing a reference to the exact data used in a study.
- **Does this data citation imply that the underlying data is publicly accessible and shared?** No.
- **Why should timestamps be used instead of semantic versioning concepts?**
Because there is no standard mechanism for determining what constitutes a 'version.'
- **How complex is it to implement the recommendations?**
It depends on the setting.
- **Why should I implement this solutions if my researchers are not asking for it or are not citing data?**
Because it's the right thing for science.

Takeaways from the paper

25

- It works and it's not as hard as it seems.
 - Not all Recommendations need to be implemented or at least not at once.
- All found value in adopting even a subset of the Recommendations because it improved services or workflows or archive practices.
- Technical migration still somewhat untested but a fact of life for archives.
- It's not really about credit.
- It's the way of the future.

- <https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html>
 - Implementation of the RDA Data Citation Recommendations by **Ocean Networks Canada (ONC)**
 - Implementation of the RDA Data Citation Recommendations the **Earth Observation Data Center (EODC) for the openEO platform**
 - **Automatically generating citation text from queries for RDBMS and XML data sources**
 - Implementing of the RDA Data Citation Recommendations by the **Climate Change Centre Austria (CCCA) for a repository of NetCDF files**
 - Implementing the RDA Data Citation Recommendations for **Long-Tail Research Data / CSV files**
 - Implementing the RDA Data Citation Recommendations in the **Distributed Infrastructure of the Virtual and Atomic Molecular Data Center (VAMDC)**
 - Implementation of Dynamic Data Citation at the **Vermont Monitoring Cooperative**
 - Adoption of the RDA Data Citation of Evolving Data Recommendation to **Electronic Health Records**

■ *Benefits*

- Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
- Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
- It allows to identify and cite even an **empty set!**
- The query stored for identifying data subsets provides valuable **provenance data**
- Query store collects **information on data usage**, offering a basis for data management decisions
- **Metadata** such as checksums support the verification of the correctness and **authenticity** of data sets retrieved
- The same principles work for **all types of data**

Agenda

28

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New” directions:
 - Challenges in complex citations
- Other issues, next steps

Any questions?

Any issues identified?

Anybody in the progress of
(planning to) implement the
recommendations?

Adoption Stories or Plans

30

- Let us know if you are (planning to) implement (part of) the recommendations
- Submit your adoption story to the RDA Webpage:

<https://www.rd-alliance.org/recommendations-outputs/adoption-stories>

Agenda

31

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New directions”:
 - Challenges in complex citations
- Other issues, next steps



Implementing the RDA WGDC Recommendations Ocean Networks Canada

Reyna Jenkins

research data sharing without barriers

rd-alliance.org



A Decade of Data 2013-2023

Research Data Alliance

exCited about Dynamic Data At Ocean Networks Canada

22 March 2023

Reyna Jenkyns, Data Stewardship Manager, ONC

ORCID: 0000-0001-6975-6816

Ocean Networks Canada, ROR: 05gknh003

University of Victoria, ROR: 04s5mat29

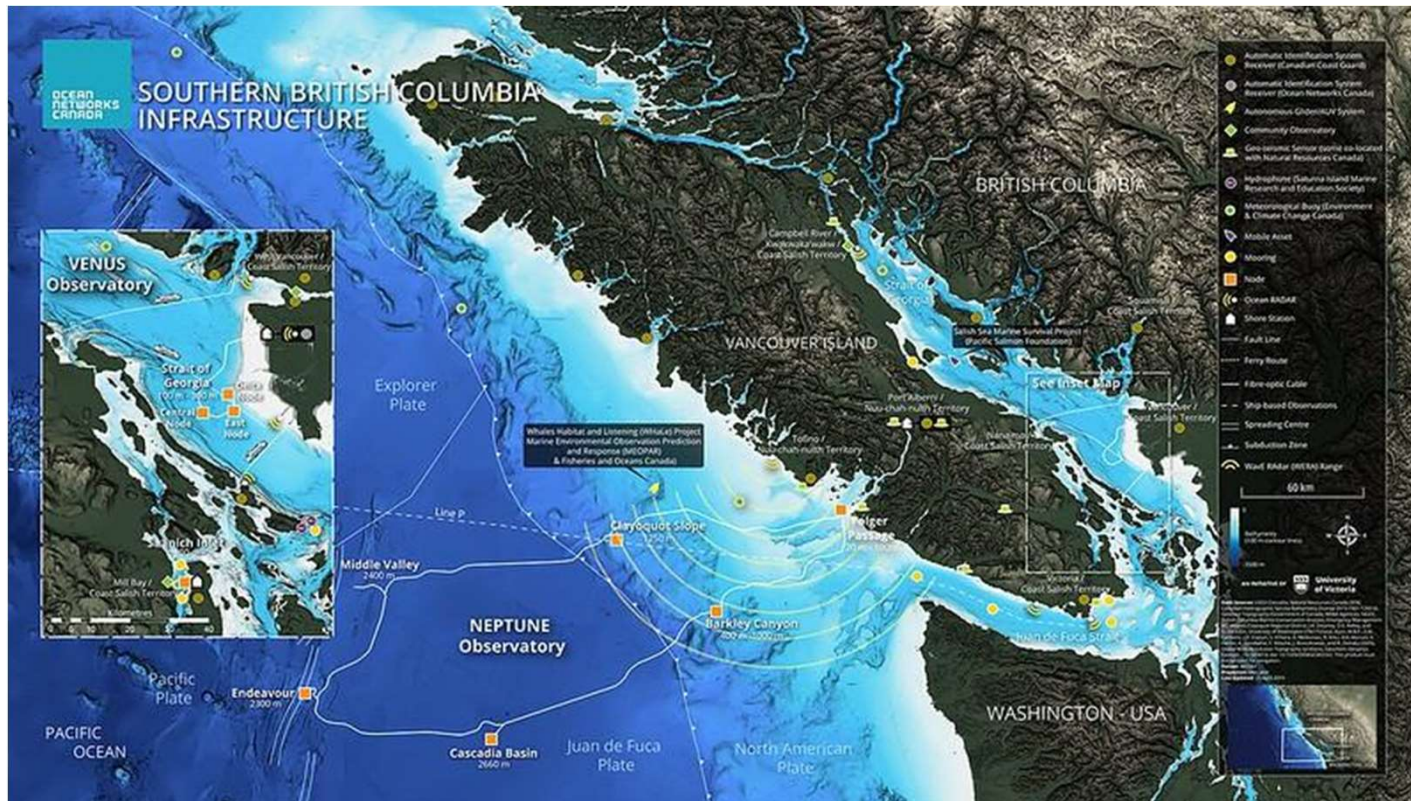
www.rd-alliance.org

Land & Sea Acknowledgement

As a University of Victoria employee, I acknowledge and respect the **ləkʷəŋən peoples** on whose traditional territory the university stands and the **Songhees, Esquimalt and W̱SÁNEĆ peoples** whose historical relationships with the land continue to this day.

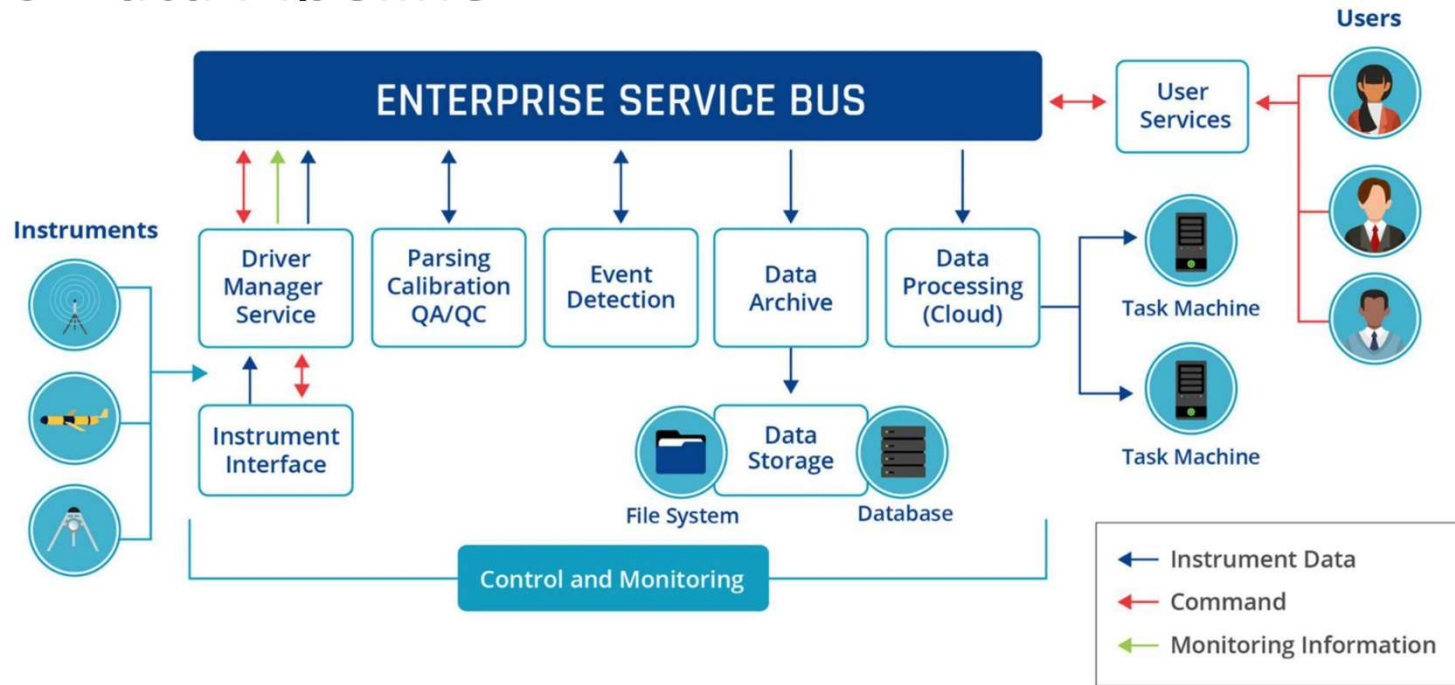
As I work at ONC, I also acknowledge the **Indigenous communities** with whom we have the honour to collaborate on **coastal monitoring** and **data management solutions**.





- Highly **heterogeneous** – fixed, mobile & profiling platforms, instrument types, data formats and processing levels, real-time vs autonomous, variety of data partners
- **Scope continually expanding** - in progress work for Biogeochemical Argo floats (about 20 bought), 6 more BC ferry routes, ,...

ONC Data Pipeline



Recent publication: Owens D, Abeysirigunawardena D, Biffard B, Chen Y, Conley P, Jenkyns R, Kerschtién S, Lavalée T, MacArthur M, Mousseau J, Old K, Paulson M, Pirenne B, Scherwath M and Thorne M (2022) **The Oceans 2.0/3.0 Data Management and Archival System**. *Front. Mar. Sci.* 9:806452. doi: 10.3389/fmars.2022.806452

Data Citation at ONC



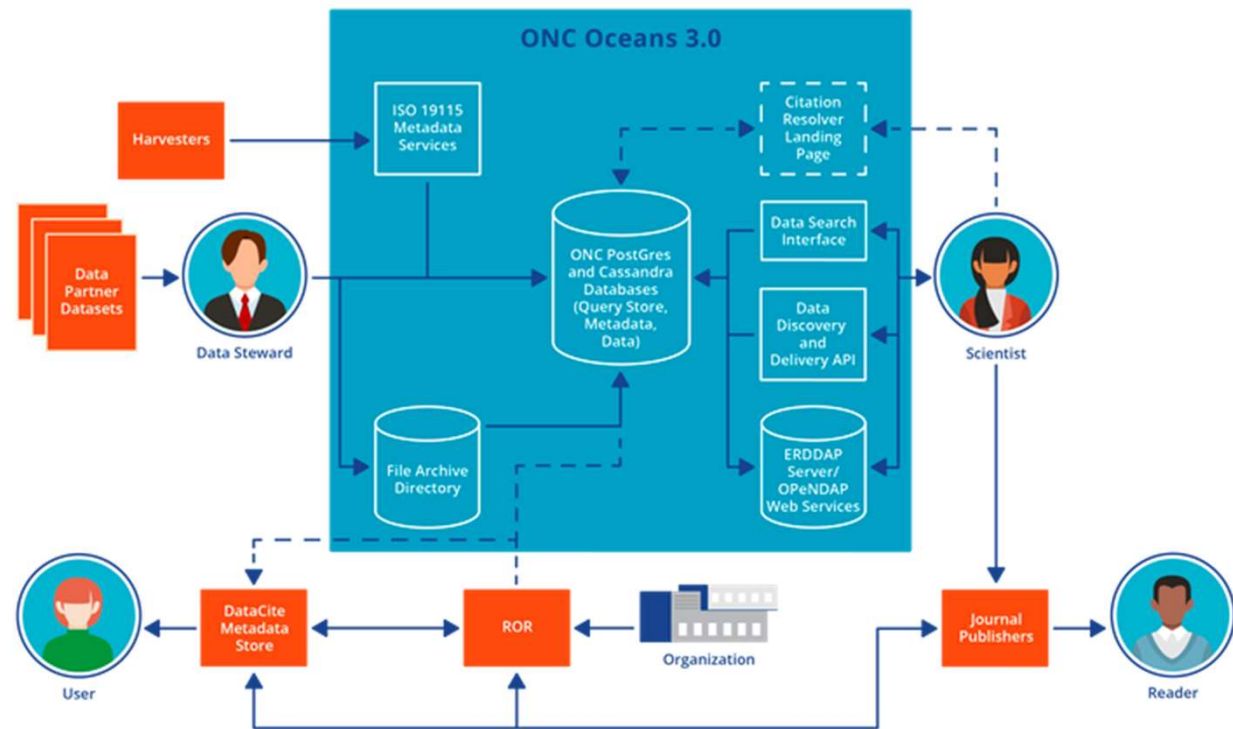
Digital Research Alliance of Canada

Alliance de recherche numérique du Canada

MINTED (Making Identifiers Necessary to Track Evolving Data), funded by Research Data Canada and CANARIE Inc. (2018-2020)

DynaCITE (Dynamic Citation Identifiers Training and Education) promotes the power of PIDs (persistent identifiers), funded by The Alliance's Data Champions Pilot Project.

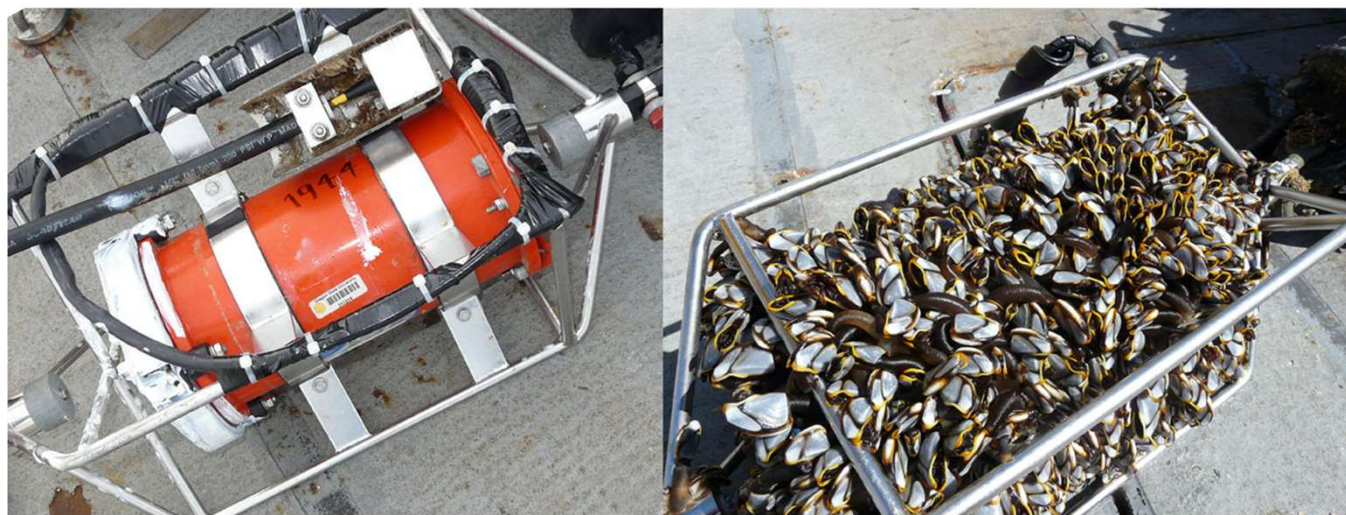
Beginning to add additional functionality and features through a project called **ExperiMINTED** for a more comprehensive citation system



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.

ONC Implementation - Automating Metadata

ONC has collected and managed thousands of datasets

Manually minting datasets DOIs would be inefficient.

Fortunately, we have a very comprehensive metadata that we can leverage for automation as the system was designed to be human/machine readable.

Automated Title

39

SearchTreeNodeName DeviceCategoryName Deployed SiteDeviceDateFrom

Upper Slope Fluorometer Turbidity Deployed 2019-05-16

Automated Abstract

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 3.0 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 3.0 digital infrastructure, with quality assurance and derived data products following established practices.

Data Search: *Subset Query Details*

[Data Source Selection](#)
Data Product Selection
[View Cart \(0 items\)](#)
[Data Search Help](#)

Fluorometer Turbidity
 WET Labs ECO FLNTUS 4670 (24117) [Details](#) | [Documentation](#)

Date From (UTC): Custom ▾

Date To (UTC): [Reset Time Fields](#)

13 Jun 2019 05 Sep 2019 28 Nov 2019 20 Feb 2020

	Time Series Scalar Data	Time Series Scalar Plot	Log File	Manual Scalar QA/QC Results
Fluorometer Turbidity 34 Annotations (Disable Pop-up Blocker to See All)	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Chlorophyll (22190)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Turbidity (22191)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Data Product Options

Quality Control: Clean Data Raw Data

Data Gaps: Fill missing/bad data with NaNs (Not a Number) Do not fill gaps

Processing:

NOTE: Most data products have additional [Metadata](#) automatically generated and added to the Cart. + Add to Cart

Query PID Landing Page: *Subset Query Details*

Ocean Networks Canada Dataset Landing Page

Oceans 2.0

Barkley Canyon instruments are offline due to node communication issue

Help | Login

Data Preview | Data Search | Plotting Utility | SeaTube | More

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
2019

Resource Type
One Deployment

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.

Data Links

User Search History (prototype)

Query PID	Device	Date From	Date To	Site ID	Dataset DOI	Copy Citation
14807859	Sea-Bird SeaCAT SBE19plus V2 7028	2017-06-27 18:02:02	2017-07-09 19:51:53	1000829	10.80242/45a5d894-046c-4b4f-b7eb-de775...	
14807859	Sea-Bird SeaCAT SBE19plus V2 7028	2017-06-27 18:02:02	2017-07-09 19:51:53	1000829	10.80242/962db9b9-eea9-4ab5-89d6-1d459...	
14807859	Sea-Bird SeaCAT SBE19plus V2 7028	2017-06-27 18:02:02	2017-07-09 19:51:53	1000829	10.80242/d792d74b-a7b-9e96-b4b7-494f7...	

43

Copy Citation

Copy Citation
 Ocean Networks Canada Society. 2017. Folger Deep Conductivity Temperature Depth Deployed 2017-05-02. Ocean Networks Canada Society. <https://doi.org/10.21383/f3299925-cf9b-4bdc-874b-d99fa0a32ef7>. Subset Query: 14807859. Accessed 2023-02-09.

CANCEL COPY TO CLIPBOARD

Versioning Data

Versioning is done in a 'batch' system

Batches contain:

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

Metadata **triggers** include::

- calibration formula revisions
- data product parameter updates
- parser modifications

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)

```
<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>
```

Version History

DOI	Reason	↓ Date Created
10.21383/5efd1457-db3f-45e0-9802-9e7e58edf004	Formula coefficient entered correctly and data needs to be reprocessed.	2020-05-10 22:34:01.414
10.21383/259fd2ac-e02d-46a0-a27b-b244b1f46dcb		2020-05-10 22:25:37.58

1-2 of 2 < 1 >

Dataset Collections

Building capacity for assigning DOIs to collections of related datasets, using the 'Collection' Resource Type of DataCite metadata

For the Collection, the DataCite metadata would indicate *hasPart* for each dataset it contains

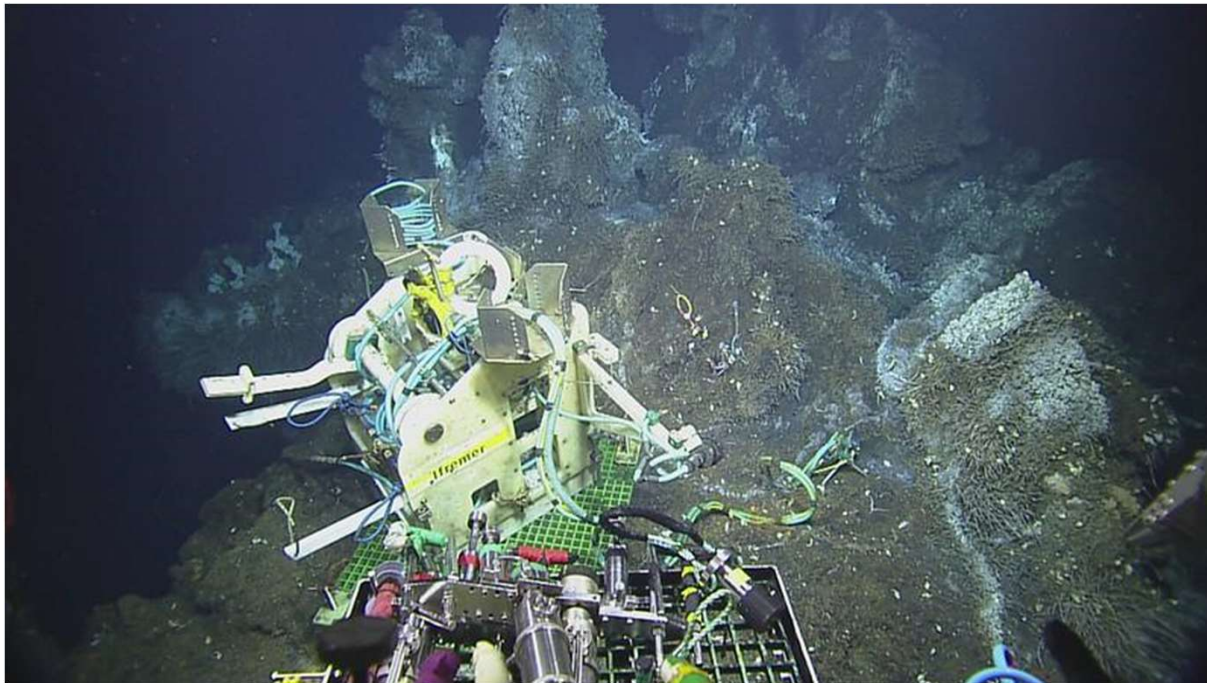
```
<relatedIdentifier
relatedIdentifierType="DOI"
relationType="HasPart"
resourceTypeGeneral="Dataset">10.3
45 34943/79343631-10ab-4a11-
b1d5-
ff55a62df097</relatedIdentifier>
```

For the Dataset, the DataCite metadata would use *isPartOf* to indicate which Collection(s) it belongs to.

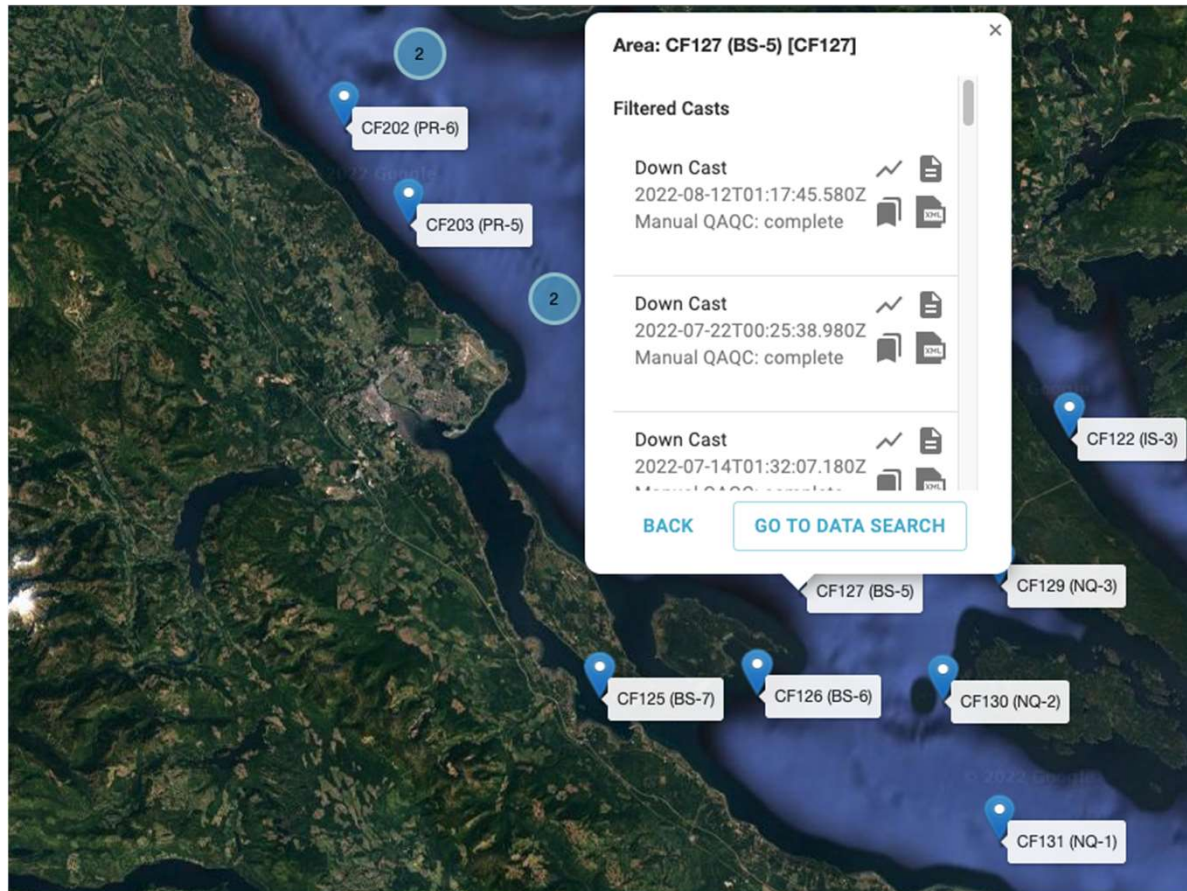
```
<relatedIdentifier
relatedIdentifierType="DOI"
relationType="isPartOf"
resourceTypeGeneral="Dataset">10.3
4943/5ab0d4d4-f8b6-478d-950b-
42fb6c6d0329</relatedIdentifier>
```


Dataset Collections - Use Cases

All datasets from a particular expedition or platform



Dataset Collections - Use Cases



Grouping of casts by station is already expressed in Oceans 3.0 metadata, but we would need to automate minting collection DOIs with DataCite and then update record for each new cast

User Requested Data Collections

Custom collections grouping data used for a particular research publication

Example:

- <https://wiki.oceannetworks.ca/display/DP/Collection%3A+Ocean+Networks+Canada+datasets+related+to+deep-water+renewal+processes+in+the+Strait+of+Georgia+from+2008-2021>
- <https://journals.ametsoc.org/view/journals/phoc/53/1/JPO-D-22-0047.1.xml>

Collection: Ocean Networks Canada datasets related to deep-water renewal processes in the Strait of Georgia from 2008-2021
Created by Chantel M Ridsdale, last modified on 23-Nov-22

Citation
Ocean Networks Canada Society. 2022. Collection: Ocean Networks Canada datasets related to deep-water renewal processes in the Strait of Georgia from 2008-2021. Ocean Networks Canada Society. <https://www.doi.org/10.26152/G315-PH34>

Title
Collection: Ocean Networks Canada datasets related to deep-water renewal processes in the Strait of Georgia from 2008-2021

DOI
10.26152/G315-PH34

JANUARY 2023 MASOUD AND PAWLOWICZ

A Predictably Intermittent Rotationally Modified Gravity Current in the Strait of Georgia

MINA MASOUD^a AND RICH PAWLOWICZ^a
^a Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia, Canada
(Manuscript received 27 February 2022, in final form 7 July 2022)

ABSTRACT: The Strait of Georgia is a large and deep fjordlike basin on the northeastern Pacific coast whose bottom waters are dramatically renewed by a series of intermittent gravity currents in summer. Here, we analyze a dataset that includes moored observations from 2008 to 2021 and shipborne measurements from a 2018 field program to describe the vertical and cross-channel structure of these gravity currents. We show that the timing of these currents for more than a decade is well predicted by proxy measurements for both tidal mixing strength in the Haro Strait/Boundary Pass region and coastal upwelling on the west coast of Vancouver Island. Renewals occur as an ~30-m-thick turbid layer extending along the right-hand slope of a broad V-shaped valley that forms the southern end of the strait. Currents are primarily along-isobath at speeds of up to 20 cm s⁻¹ with a small downhill component. A diagnostic analytical model with a depth-dependent eddy viscosity is fitted to the observations and confirms a clockwise rotation of current vectors with height, partly driven by boundary layer dynamics over a scale of a few meters and partly driven by Coriolis forces in the near-bottom linear density gradient. Bottom drag and (small) entrainment parameters are similar to those found in other oceanic situations, and the current is “laminar” with respect to large-scale instabilities (with Froude number ~1 and Ekman number ~0.01), although subject to turbulence at small scales (Reynolds number of ~10⁶). The predictability and reliability of this accessible rotationally modified gravity current suggests that it is an ideal geophysical laboratory for future studies of such features.

48

Ocean Networks Canada Society. 2022. Collection: Ocean Networks Canada datasets related to deep-water renewal processes in the Strait of Georgia from 2008-2021. Ocean Networks Canada Society. <https://www.doi.org/10.26152/G315-PH34>

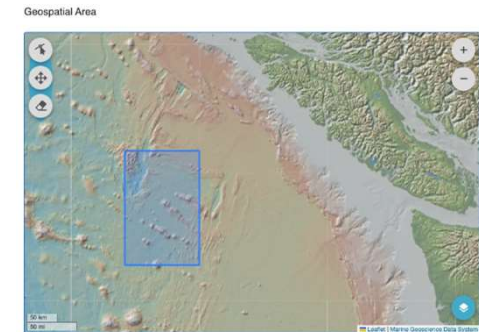
ExperimINTED

This project has 8 goals:

- **Goal 1:** Make the DOI/query PID information more obvious to end-users in Oceans 3.0
 - One mechanism is Search History prototype in earlier slide
- **Goal 2:** Make the Batch Versioning System more robust and comprehensive in the use cases it can support.
- **Goal 3:** Ensure versioning provenance of data is more fully represented and adhering to best practices.
 - integrate dataset versioning information into ISO 19115 metadata records
 - represent versioning changes using the W3C PROV standard (see <https://www.w3.org/TR/prov-overview/>)

ExperimINTED

- **Goal 4:** Enhance the Dataset Landing pages to have richer metadata, relationships, and discoverability.
 - schema.org integration to enable Google search
 - map for geoextent of dataset
- **Goal 5:** Leverage more of the DataCite Metadata Schema for improved accuracy and relationships, especially publication relationships
- **Goal 6:** Introduce metrics for tracking and reporting on data citations, leveraging PID graph
- **Goal 7:** Introduce dataset collections that can be automated for common use-cases, and manually generated for user-specific cases.
- **Goal 8:** Ensure that the data citations and related infrastructure are monitored and preserved.





OCEAN
NETWORKS
CANADA

Questions?
reyna@oceannetworks.ca

Agenda

52

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New directions”:
 - Challenges in complex citations
- Other issues, next steps



RDA WGDC Recommendations Challenges for Complex Citations?

Deb Agarwal, Shelly Stall

research data sharing without barriers

rd-alliance.org

Why a Complex Citation WG? Isn't this already covered?

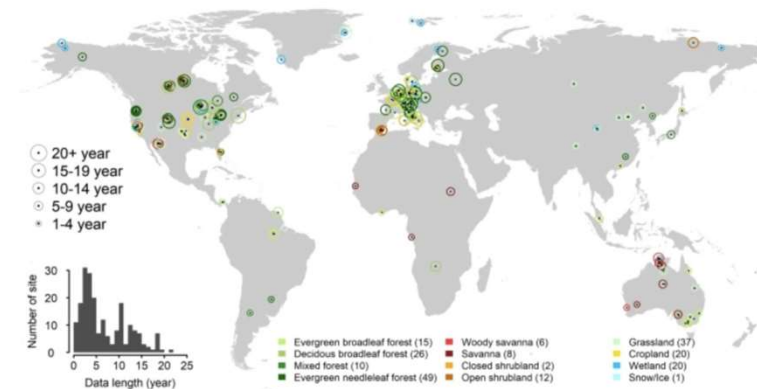
Deb Agarwal, Shelley Stahl , Lesley Wyburn, Martina Stockhause

Challenge

- Need to cite > 50 (could be millions) of resources in a paper
- The resources come from many different repositories
- The usage license requires acknowledgement by citation (e.g. CC By 4.0 where citation is the acknowledgement)
- Current typical publisher guidance to paper authors (even for data papers)
 - Add citations to the supplementary materials
 - Cons – not machine readable and unclear that this meets the usage license
 - Drop, combine, or summarize citations

Simplest Case - Global FLUXNET 2015 set is made of up of 212 datasets worldwide

Enabling citation - First ideas



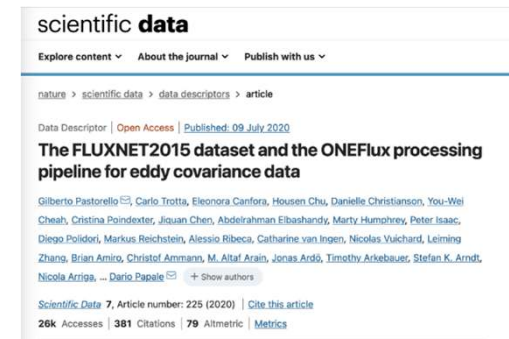
- Collection – users cite the collection
 - Pro – the FLUXNET repository could implement (but not trivial)
 - Con – no reusability/traceability of what data was actually used
 - Con – extremely difficult to trace credit to datasets
 - Con - no authorship in the citation
- Dynamic data citation – user gets a custom citation for their query
 - Con - no authorship in the citation
 - Con – not currently implemented by FLUXNET repository nor is tracing of the credit
 - Con – inaccurate citation since ~50% of the users download the whole FLUXNET set and then actually use a subset

Simplest case – final solution

- Write a high-prestige paper about the set of datasets that can serve as the citation for the set
 - Include all personnel involved in the datasets as authors (limited to 500 by publisher – only a few contributors from each dataset)
 - Include the central team that processes and helps to build the set as authors
- Implementation of paper
 - Needed to get agreement from all datasets (some never agreed)
 - Publisher required that the set be substantially different than the set already released (required collection of significant additional data)
 - Required development of new meta analyses, graphs, and conclusions related to the data.
 - Required agreement of CC BY 4.0 as data usage policy and author order
 - Took 4 years to produce and publish

Paper result

- Pros
 - Highly cited paper - 508 citations (Google scholar)
 - Authors on the paper have not complained
 - Users are able to utilize non-trivial subsets in analyses and cite the data
 - Central processing team were included in the writing of the paper
- Cons
 - The authors included only the senior contributors to each dataset (PIs)
 - All datasets treated as equally likely to be used when they are not
 - Limited reusability/traceability of which data used (paper may contain a table listing sites)
 - Moving to regular releases of FLUXNET – not realistic to write a paper for each release



Pastorello, G., Trotta, C., Canfora, E. et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data* 7, 225 (2020). <https://doi.org/10.1038/s41597-020-0534-3>

Additional use cases

- UN IPCC climate reports – need to be able to cite a large number of resources used to produce a figure, traceability essential for the transparency of IPCC's findings
- Oceanographic data – repository is building new data products consisting of parts of many stored datasets – using dynamic citation but need additional features
- ... still collecting use cases

Complex Citation

- Need a solution that enables > 50 citations for resources from across many repositories
- Need to be able to trace paper citations to the underlying resource
- Community of Practice has been meeting since Dec. 2020
 - Gathered use cases (many more coming in)
 - Includes publishers, repositories, data repositories, ...
 - Refer to the needed result as reliquary (avoid name collision)
 - Proof of concept 'cocktail napkin' that works for multiple use cases
- Hoping a current mechanism can be leveraged to meet the complex citation needs of a reliquary

Agenda

61

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Update on Adoptions:
 - Ocean Network Canada
- “New directions”:
 - Challenges in complex citations
- Other issues, next steps

Thanks

62

Thanks!
And hope to see you at the
next meeting
of the
WGDC