DataCite

CONNECTING RESEARCH,
ADVANCING KNOWLEDGE

DataCite Training:
Citations in DataCite Metadata

Kelly Stathis, Iratxe Puebla, Mary Hirsch

12th of October 2023
DataCite Community Meeting

@datacite

@datacite@openbiblio.social

Learning objectives

This training is designed for DataCite Members, Consortium Organizations and Repositories.

In this training you will learn everything you need to know about how citation works from DataCite's perspective. This includes: the actions you can take to **update DataCite DOI metadata** to create citations when your resources are cited and what happens to the metadata to **make citation counts for your resources visible to the whole community**.

What we'll cover

Part 1: What is citation and why does citation matter?

Part 2: Adding citations to DataCite metadata

Part 3: Using citations in DataCite metadata

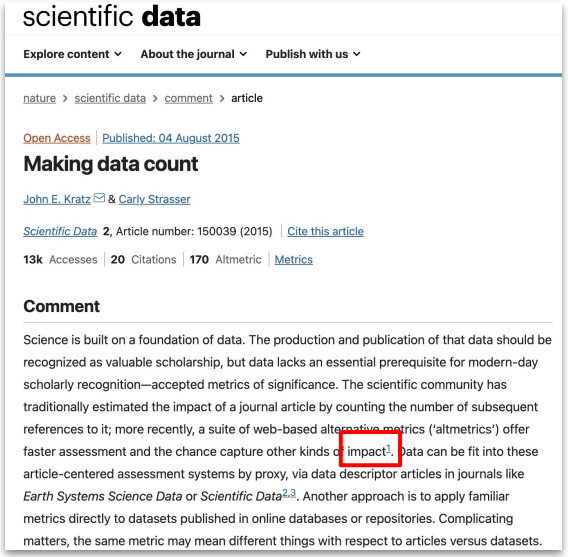
Part 4: The Global Data Citation Corpus

Part 1: What is citation and why does it matter?

Citations

"A citation is a formal structured reference to another scholarly published or unpublished work" – FORCE11 Data Citation principles

As part of publications, sources are generally mentioned in the text (marked or in abbreviated format) and then included in a bibliography or references section.



scientific **data**

Explore content | About the journal | Publish with us

nature › scientific data › comment › article

Open Access | Published: 04 August 2015

Making data count

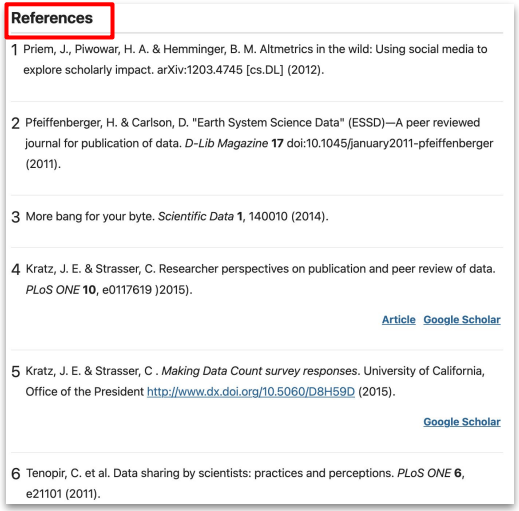
John E. Kratz & Carly Strasser

Scientific Data **2**, Article number: 150039 (2015) | Cite this article

13k Accesses | **20** Citations | **170** Altmetric | Metrics

Comment

Science is built on a foundation of data. The production and publication of that data should be recognized as valuable scholarship, but data lacks an essential prerequisite for modern-day scholarly recognition—accepted metrics of significance. The scientific community has traditionally estimated the impact of a journal article by counting the number of subsequent references to it; more recently, a suite of web-based alternative metrics ('altmetrics') offer faster assessment and the chance capture other kinds of impact[1]. Data can be fit into these article-centered assessment systems by proxy, via data descriptor articles in journals like *Earth Systems Science Data* or *Scientific Data*[2,3]. Another approach is to apply familiar metrics directly to datasets published in online databases or repositories. Complicating matters, the same metric may mean different things with respect to articles versus datasets.



References

1 Priem, J., Piwowar, H. A. & Hemminger, B. M. Altmetrics in the wild: Using social media to explore scholarly impact. arXiv:1203.4745 [cs.DL] (2012).

2 Pfeiffenberger, H. & Carlson, D. "Earth System Science Data" (ESSD)—A peer reviewed journal for publication of data. *D-Lib Magazine* **17** doi:10.1045/january2011-pfeiffenberger (2011).

3 More bang for your byte. *Scientific Data* **1**, 140010 (2014).

4 Kratz, J. E. & Strasser, C. Researcher perspectives on publication and peer review of data. *PLoS ONE* **10**, e0117619)2015).

Article Google Scholar

5 Kratz, J. E. & Strasser, C . *Making Data Count survey responses*. University of California, Office of the President http://www.dx.doi.org/10.5060/D8H59D (2015).

Google Scholar

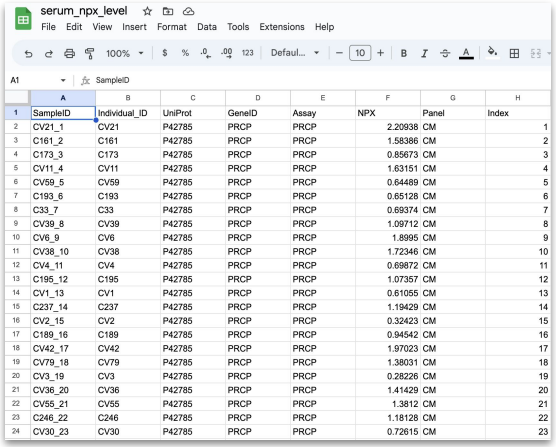
6 Tenopir, C. et al. Data sharing by scientists: practices and perceptions. *PLoS ONE* **6**, e21101 (2011).

https://doi.org/10.1038/sdata.2015.39

Citations

Citation means reference to another object the researchers have used as part of their work, this may be articles, data, software or other objects.

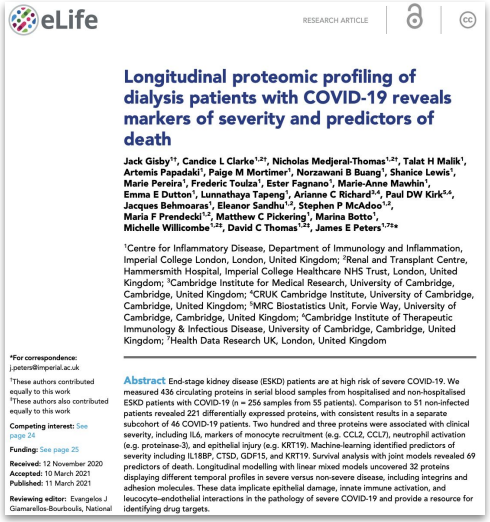
Citations are useful as they clearly point to the object being used or reused in research - they provide a link between the research objects


serum_npx_level

File Edit View Insert Format Data Tools Extensions Help

	A	B	C	D	E	F	G	H
1	SampleID	Individual_ID	UniProt	GeneID	Assay	NPX	Panel	Index
2	CV21_1	CV21	P42785	PRCP	PRCP	2.20938	CM	1
3	C161_2	C161	P42785	PRCP	PRCP	1.58386	CM	2
4	C173_3	C173	P42785	PRCP	PRCP	0.85673	CM	3
5	CV11_4	CV11	P42785	PRCP	PRCP	1.63151	CM	4
6	CV59_5	CV59	P42785	PRCP	PRCP	0.64489	CM	5
7	C193_6	C193	P42785	PRCP	PRCP	0.65128	CM	6
8	C33_7	C33	P42785	PRCP	PRCP	0.69374	CM	7
9	CV39_8	CV39	P42785	PRCP	PRCP	1.09712	CM	8
10	CV6_9	CV6	P42785	PRCP	PRCP	1.8995	CM	9
11	CV38_10	CV38	P42785	PRCP	PRCP	1.72346	CM	10
12	CV4_11	CV4	P42785	PRCP	PRCP	0.69872	CM	11
13	C195_12	C195	P42785	PRCP	PRCP	1.07357	CM	12
14	CV1_13	CV1	P42785	PRCP	PRCP	0.61055	CM	13
15	C237_14	C237	P42785	PRCP	PRCP	1.19429	CM	14
16	CV2_15	CV2	P42785	PRCP	PRCP	0.32423	CM	15
17	C189_16	C189	P42785	PRCP	PRCP	0.94542	CM	16
18	CV42_17	CV42	P42785	PRCP	PRCP	1.97023	CM	17
19	CV79_18	CV79	P42785	PRCP	PRCP	1.38031	CM	18
20	CV3_19	CV3	P42785	PRCP	PRCP	0.28226	CM	19
21	CV36_20	CV36	P42785	PRCP	PRCP	1.41429	CM	20
22	CV55_21	CV55	P42785	PRCP	PRCP	1.3812	CM	21
23	C246_22	C246	P42785	PRCP	PRCP	1.18128	CM	22
24	CV30_23	CV30	P42785	PRCP	PRCP	0.72615	CM	23

Gisby J, Clarke CL, Medjeral-Thomas N, Malik TH, Papadaki A, Mortimer PM, Buang NB, Lewis S, Pereira M, Toulza F, Fagnano E, Mawhin M, Dutton EE, Tapeng L, Kirk P, Behmoaras J, Sandhu E, McAdoo SP, Prendecki MF, Pickering MC, Botto M, Willicombe W, Thomas DC, Peters JE (2020) **Dryad Digital Repository** Longitudinal proteomic profiling of high-risk patients with COVID-19 reveals markers of severity and predictors of fatal disease. https://doi.org/10.5061/dryad.6t1g1jwxj


eLife

RESEARCH ARTICLE

Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death

Jack Gisby[†1], Candice L Clarke[†1,2†], Nicholas Medjeral-Thomas[†1,2†], Talat H Malik[1], Artemis Papadaki[1], Paige M Mortimer[1], Norzawani B Buang[1], Shanice Lewis[1], Marie Pereira[1], Frederic Toulza[1], Ester Fagnano[1], Marie-Anne Mawhin[1], Emma E Dutton[1], Lunnathaya Tapeng[1], Arianne C Richard[3,4], Paul DW Kirk[5,6], Jacques Behmoaras[1], Eleanor Sandhu[1,2], Stephen P McAdoo[1,2], Maria F Prendecki[1,2], Matthew C Pickering[1], Marina Botto[1], Michelle Willicombe[1,2‡], David C Thomas[1,2‡], James E Peters[1,7‡*]

[1]Centre for Inflammatory Disease, Department of Immunology and Inflammation, Imperial College London, London, United Kingdom; [2]Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, United Kingdom; [3]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom; [4]CRUK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; [5]MRC Biostatistics Unit, Forvie Way, University of Cambridge, Cambridge, United Kingdom; [6]Cambridge Institute of Therapeutic Immunology & Infectious Disease, University of Cambridge, Cambridge, United Kingdom; [7]Health Data Research UK, London, United Kingdom

*For correspondence: j.peters@imperial.ac.uk

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Competing interest: See page 24

Funding: See page 25

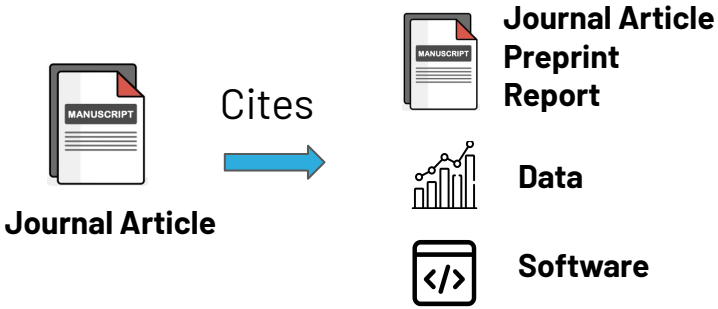
Received: 12 November 2020
Accepted: 10 March 2021
Published: 11 March 2021

Reviewing editor: Evangelos J Giamarellos-Bourboulis, National

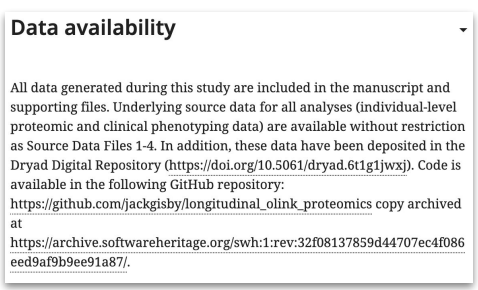
Abstract End-stage kidney disease (ESKD) patients are at high risk of severe COVID-19. We measured 436 circulating proteins in serial blood samples from hospitalised and non-hospitalised ESKD patients with COVID-19 (n = 256 samples from 55 patients). Comparison to 51 non-infected patients revealed 221 differentially expressed proteins, with consistent results in a separate subcohort of 46 COVID-19 patients. Two hundred and three proteins were associated with clinical severity, including IL6, markers of monocyte recruitment (e.g. CCL2, CCL7), neutrophil activation (e.g. proteinase-3), and epithelial injury (e.g. KRT19). Machine-learning identified predictors of severity including IL18BP, CTSD, GDF15, and KRT19. Survival analysis with joint models revealed 69 predictors of death. Longitudinal modelling with linear mixed models uncovered 32 proteins displaying different temporal profiles in severe versus non-severe disease, including integrins and adhesion molecules. These data implicate epithelial damage, innate immune activation, and leucocyte–endothelial interactions in the pathology of severe COVID-19 and provide a resource for identifying drug targets.

Citations

Most commonly citations happen when researchers cite scholarly objects in articles.

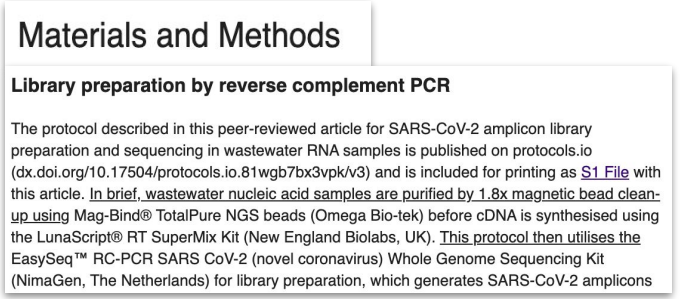


Something to bear in mind in the context of journal articles is that citations can appear in different locations in the article.



Data availability

All data generated during this study are included in the manuscript and supporting files. Underlying source data for all analyses (individual-level proteomic and clinical phenotyping data) are available without restriction as Source Data Files 1-4. In addition, these data have been deposited in the Dryad Digital Repository (https://doi.org/10.5061/dryad.6t1g1jwxj). Code is available in the following GitHub repository: https://github.com/jackgisby/longitudinal_olink_proteomics copy archived at https://archive.softwareheritage.org/swh:1:rev:32f08137859d44707ec4f086eed9af9b9ee91a87/.



Materials and Methods

Library preparation by reverse complement PCR

The protocol described in this peer-reviewed article for SARS-CoV-2 amplicon library preparation and sequencing in wastewater RNA samples is published on protocols.io (dx.doi.org/10.17504/protocols.io.81wgb7bx3vpk/v3) and is included for printing as S1 File with this article. In brief, wastewater nucleic acid samples are purified by 1.8x magnetic bead clean-up using Mag-Bind® TotalPure NGS beads (Omega Bio-tek) before cDNA is synthesised using the LunaScript® RT SuperMix Kit (New England Biolabs, UK). This protocol then utilises the EasySeq™ RC-PCR SARS CoV-2 (novel coronavirus) Whole Genome Sequencing Kit (NimaGen, The Netherlands) for library preparation, which generates SARS-CoV-2 amplicons

References

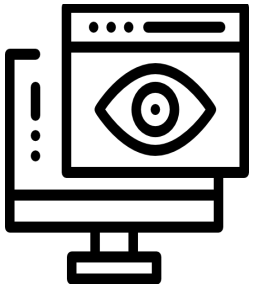
Abdill, R. J. and Blekhman, R. (2019). Meta-research: tracking the popularity and outcomes of all bioRxiv preprints. *ELife* 8, e45133. https://doi.org/10.7554/eLife.45133

Dataset and softwareicons by Freepik and Arkinasi via Flaticon, manuscript icon by ASAPbio

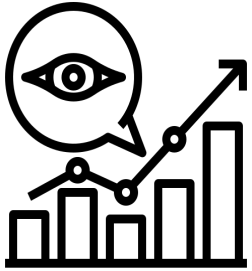
The benefits of Citations

Acknowledgement for the researchers who produce the object

Transparency & reproducibility for those using the object



Visibility for the repositories that host research objects



Evaluation of open outputs

Icons by Freepik and ultimatearn via Flaticon

Part 2: Adding citations to DataCite metadata

Part 2: Adding citations in DataCite metadata

- How do I include citations and references in DOI metadata using Related Identifiers?
- What counts as a citation? What counts as a reference?
- What happens to RelatedIdentifiers that I add to my DOIs?

DataCite DOI metadata

In this section we are going to be using a lot of terminology from the DataCite metadata schema.

If you are not familiar with the schema, we highly recommend you watch our metadata trainings to help give you more context:

- DataCite Metadata Training: (Part 1 - the metadata schema) (https://youtu.be/YY4BUXO0-q4)

- DataCite Metadata Training: (Part 2 - connection metadata) (https://youtu.be/P70hbhrbWPA)

DataCite DOI metadata

We will focus on including citations in DataCite DOI metadata from the repository perspective.

This is applicable to all resources with DataCite DOIs, including:

- Datasets
- Software
- Images
- Preprints
- etc.

DataCite Metadata Schema
20 metadata properties

6 mandatory:

- Identifier, Creator, Title, Publisher, PublicationYear, ResourceType

6 recommended:

- Subject, Contributor, Date, RelatedIdentifier, Description, GeoLocation

8 optional:

- Language, AlternateIdentifier, Size, Format, Version, Rights, FundingReference, RelatedItem

Related Identifiers
How we make connections

RelatedIdentifier is a property in the DataCite Metadata Schema used to connect research outputs to other research outputs.

Every relatedIdentifier has a **relationType** which defines the type of relationship, for example:

- DOI A *HasPart* DOI B (whole/part relationship)
- DOI A *IsCitedBy* DOI B (citation relationship)

relationType pairs
Reciprocal relationships

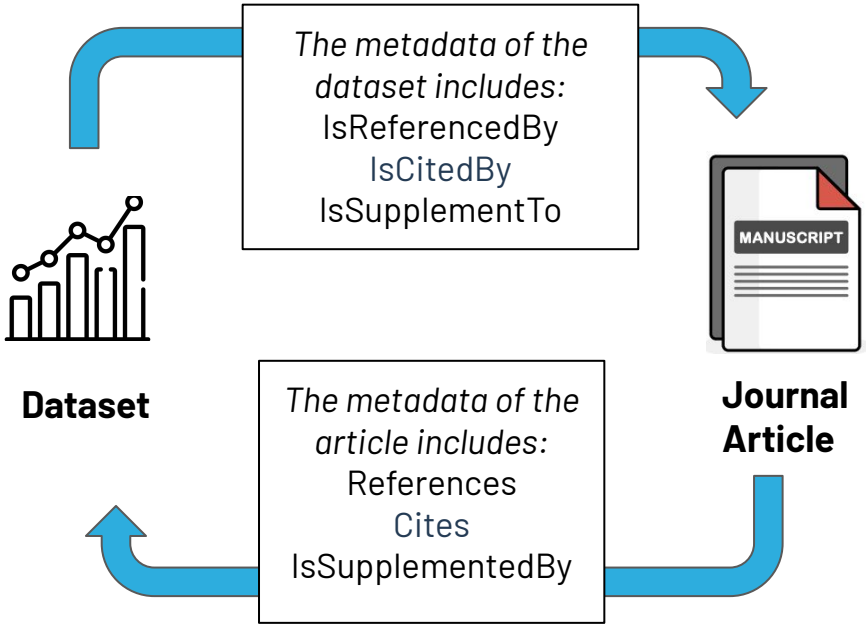
Most relationTypes work in reciprocal pairs, e.g.:

- *HasPart* and *IsPartOf*
- *IsCitedBy* and *Cites*

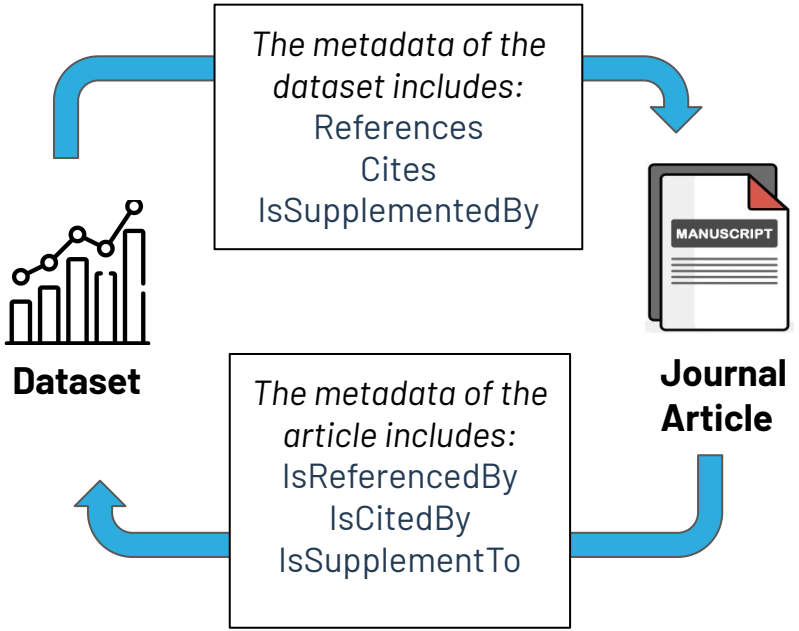
For example:

- If DOI A *HasPart* DOI B, you can also say that DOI B *IsPartOf* DOI A.
- If DOI A *IsCitedBy* DOI B, you can also say that DOI B *Cites* DOI A.

How to create citations



Dataset citation



Journal Article citation

Dataset icons by Freepik via Flaticon, manuscript icon by ASAPbio

https://support.datacite.org/docs/contributing-citations-and-references

Citations and references
3 relationType pairs

DataCite

For citations and references, there are only three relationType pairs that count:

IsCitedBy and **Cites**

IsReferencedBy and **References**

IsSupplementTo and **IsSupplementedBy**

Example 1: A dataset has been cited in a journal article

A data repository hosts a **dataset A** that is cited by a **journal article B**.

When registering the DOI for dataset A, they can include a relatedIdentifier for paper B with any of the following relationTypes:

- IsCitedBy
  - "Dataset A is cited by journal article B"
- IsReferencedBy
  - "Dataset A is referenced by journal article B"
- IsSupplementTo
  - "Dataset A is supplement to journal article B"

This is interpreted as a **citation** for dataset A, or a **reference** from journal article B.

Example 1: A dataset has been cited in a journal article

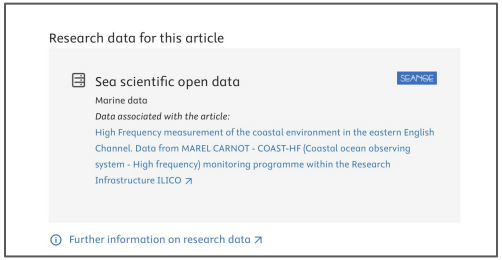
[DATASET] with a DataCite DOI:

High Frequency measurement of the coastal environment in the eastern English Channel. Data from MAREL CARNOT - COAST-HF (Coastal ocean observing system - High frequency) monitoring programme within the Research Infrastructure ILICO

IsCitedBy

[JOURNAL ARTICLE] with a Crossref DOI:

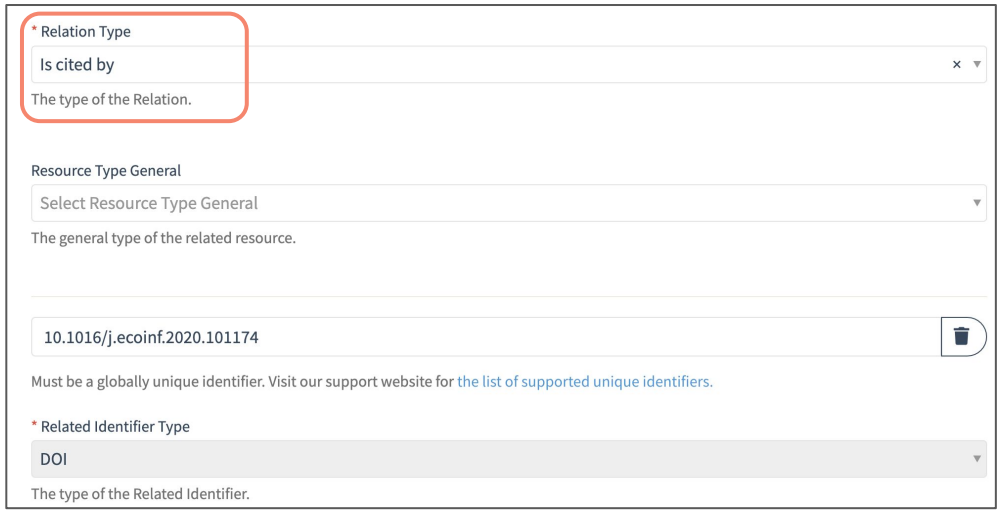
Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas



relationType metadata

DataCite

DATACITE FABRICA FORM



DATACITE XML

```
<relatedIdentifier
relatedIdentifierType="DOI"
relationType="IsCitedBy">10.1016/j.ecoinf
.2020.101174</relatedIdentifier>
```

DATACITE JSON

```
{
    "relationType": "IsCitedBy",
    "relatedIdentifier":
"10.1016/j.ecoinf.2020.101174",
    "relatedIdentifierType": "DOI"
},
```

Example 2: A dissertation cites a dataset

DataCite

An institutional repository hosts a **dissertation C** which cites a **dataset D.**

When registering the DOI for dissertation C, they can include a RelatedIdentifier for dataset D with any of the following relationTypes:

- Cites
  - “Dissertation C cites dataset D”
- References
  - “Dissertation C references dataset D”
- IsSupplementedBy
  - “Dissertation C is supplemented by dataset D”

This is interpreted as a **citation** for dataset D, or a **reference** from dissertation C.

Example 2: A dissertation cites a dataset

[DISSERTATION] with a DataCite DOI:
Research data management

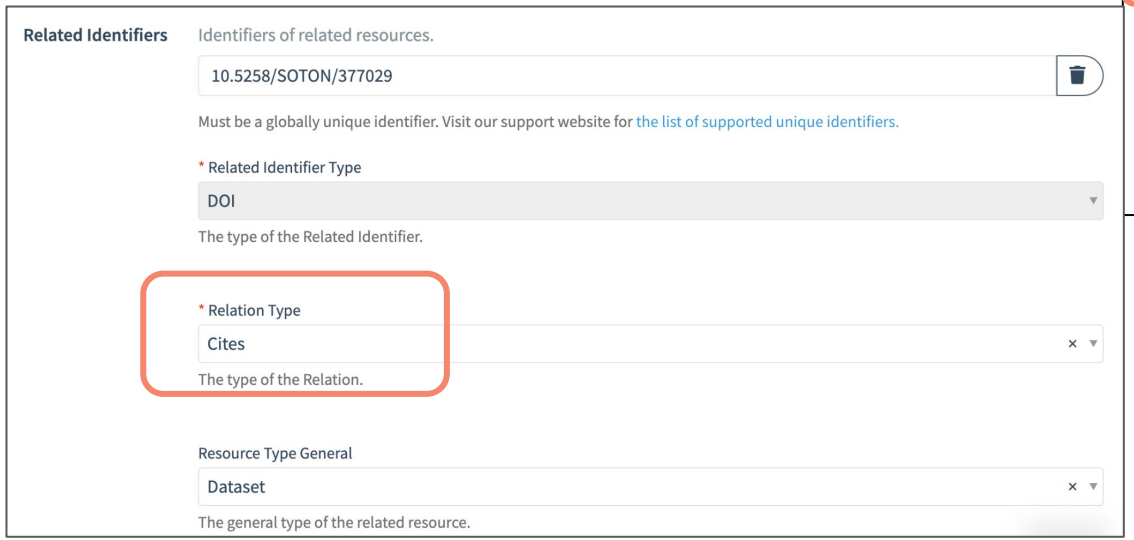
Cites

[DATASET] with a DataCite DOI:
Heterogeneous Data Centre CodePlex Repository.

relationType metadata

DataCite

DATACITE FABRICA FORM



DATACITE XML

```
<relatedIdentifiers>

        <relatedIdentifier relatedIdentifierType="DOI"
relationType="Cites"
resourceTypeGeneral="Dataset">10.5258/SOTON/377029</related
Identifier>

    </relatedIdentifiers>
```

DATACITE JSON

```
"relatedIdentifiers": [
    {
      "schemeUri": null,
      "schemeType": null,
      "relationType": "Cites",
      "relatedIdentifier":
"10.5258/SOTON/377029",
      "resourceTypeGeneral": "Dataset",
      "relatedIdentifierType": "DOI",
      "relatedMetadataScheme": null
    }
  ],
```

What you should include as a repository

When you are updating metadata for a given DOI, remember:

- To add a **citation for the DOI you are updating**, use IsCitedBy, IsReferencedBy, or IsSupplementTo.

- To **cite another DOI**, use Cites, References, or IsSupplementedBy.

Part 3: Using citations in DataCite metadata

What happens to RelatedIdentifiers?

RelatedIdentifiers for DOIs and URLs are included in **DataCite Event Data**, which is a store of connections ("linking events") that are generated when PIDs are connected through metadata.

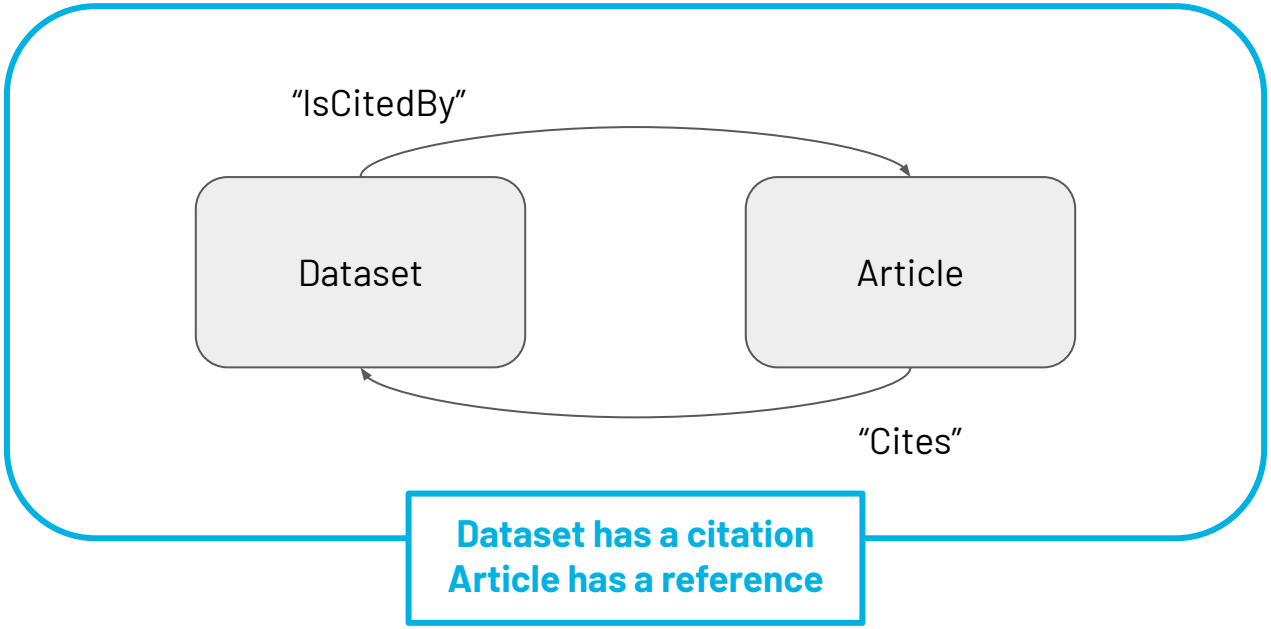
DataCite Event Data underlies several DataCite services, including:

- DataCite Commons
- ORCID auto-update
- Global Data Citation Corpus (in development)

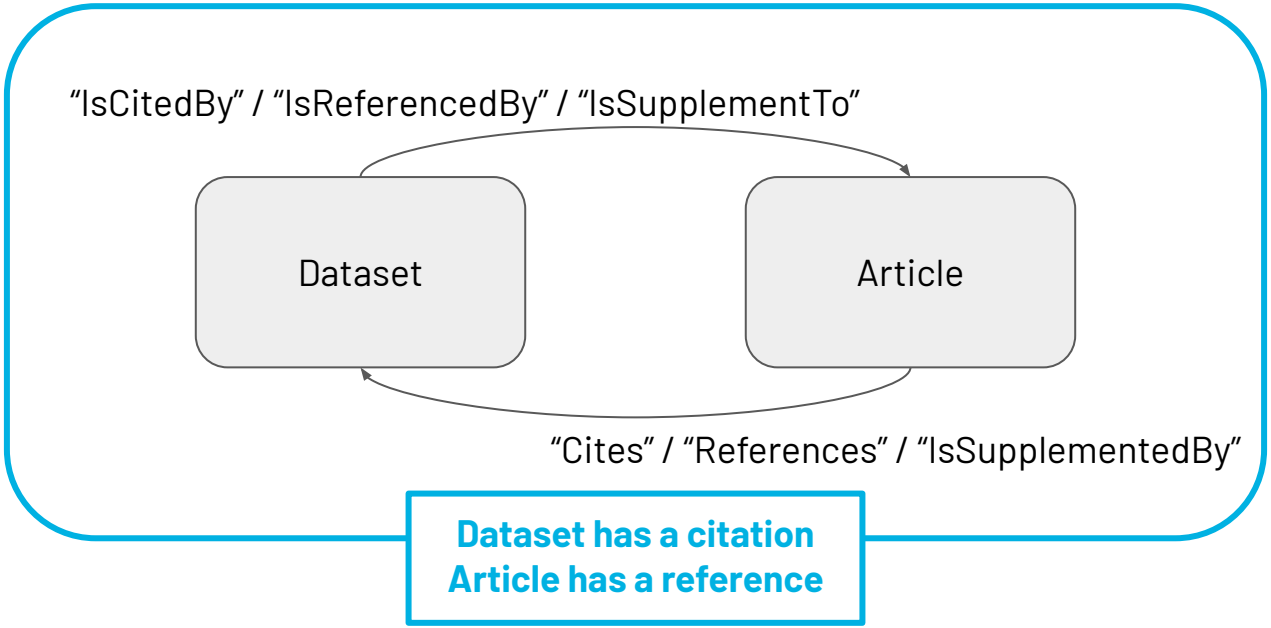
Which RelatedIdentifiers are included as citations and references?

- Must link to another DOI
- Must have one of the relationTypes:
  - IsCitedBy / Cites
  - IsReferencedBy / References
  - IsSupplementTo / IsSupplementedBy

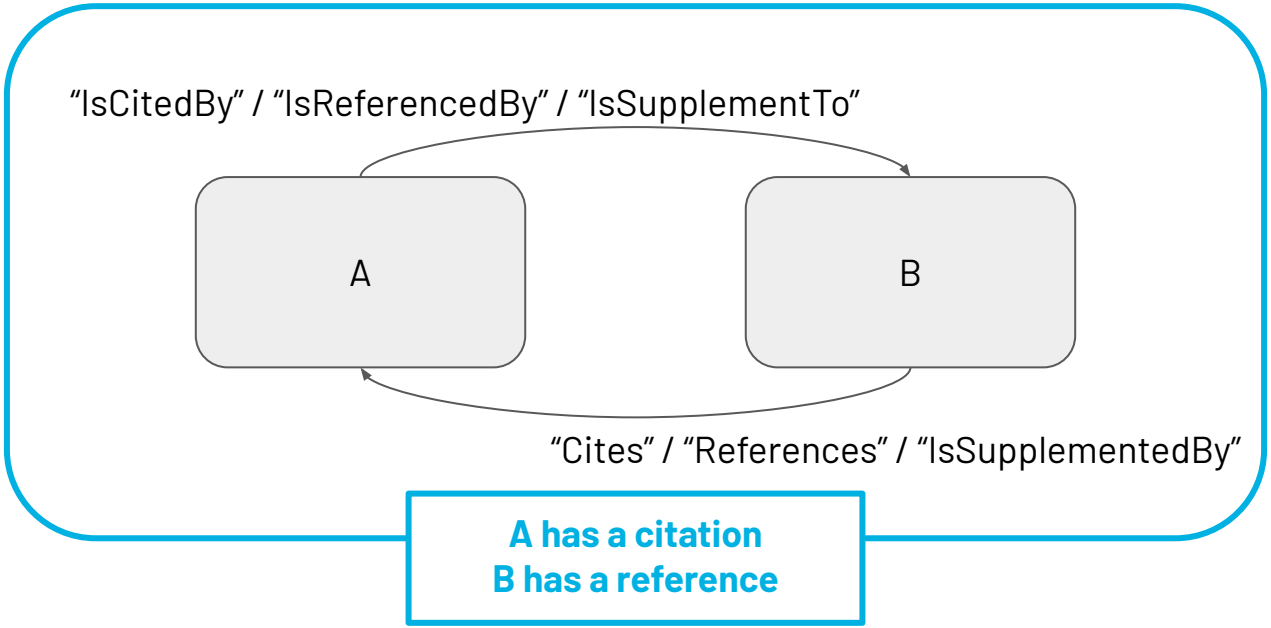
How DataCite interprets relationTypes



How DataCite interprets relationTypes



How DataCite interprets relationTypes



How DataCite interprets relationTypes

DataCite

relationType in metadata for DOI "A"	Relationship between A and B	Equivalent to	Counts as a citation for	Counts as a reference for
IsCitedBy	A is cited by B	B cites A	A	B
IsReferencedBy	A is referenced by B	B references A	A	B
IsSupplementTo	A is supplement to B	B is supplemented by A	A	B
Cites	A cites B	B is cited by A	B	A
References	A references B	B is referenced by A	B	A
IsSupplementedBy	A is supplemented by B	B is supplement to A	B	A

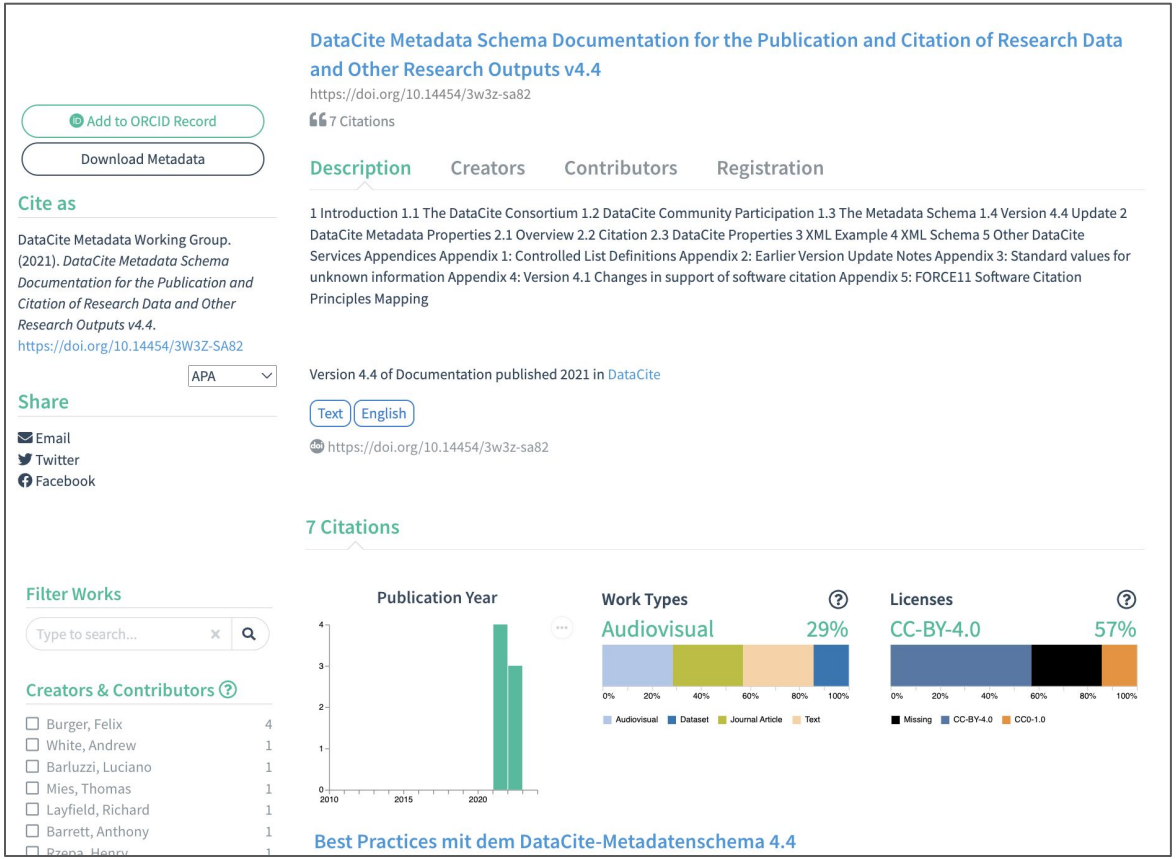
https://support.datacite.org/docs/contributing-citations-and-references

Part 3: Using citations in DataCite metadata

There are several ways to retrieve data citations from DataCite:

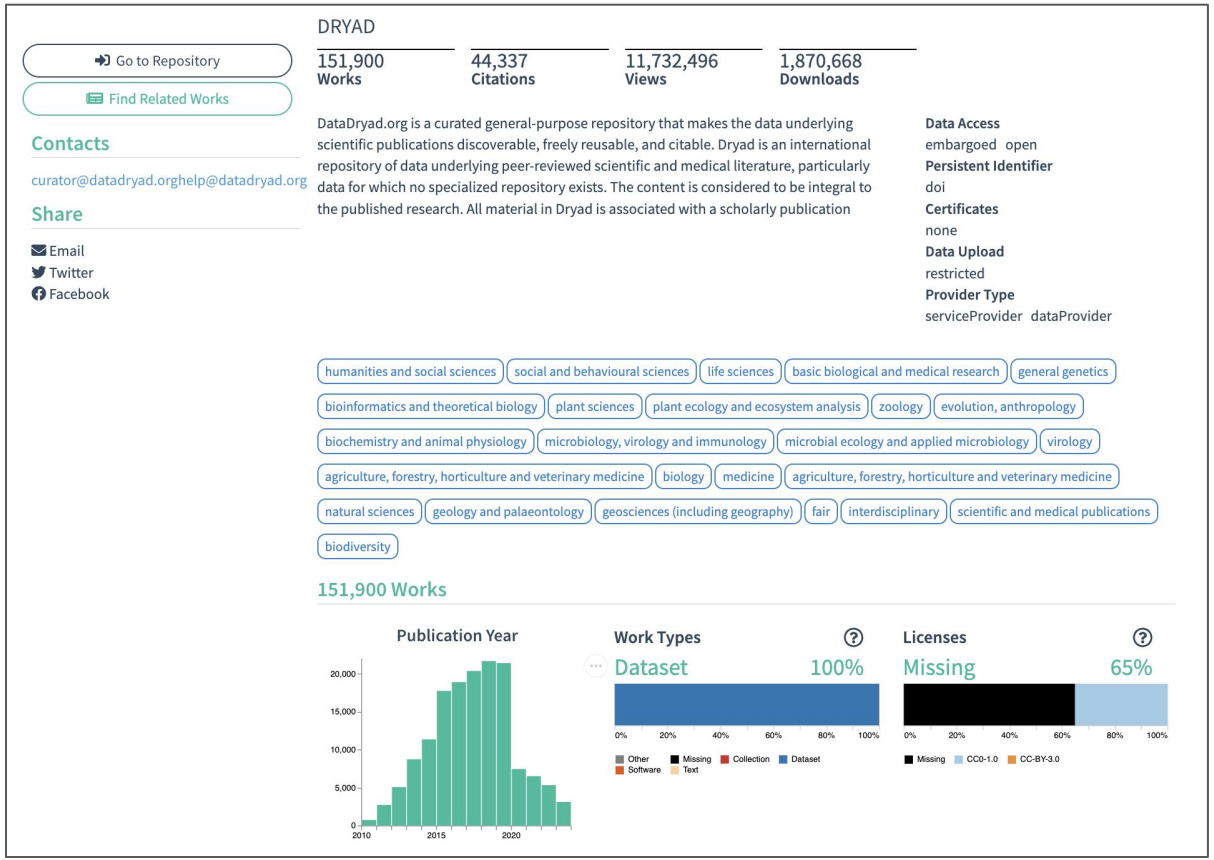
- DataCite Commons
- DataCite REST API
- Next steps for the Global Data Citation Corpus

DataCite Commons: Work



https://commons.datacite.org/doi.org/10.14454/3w3z-sa82

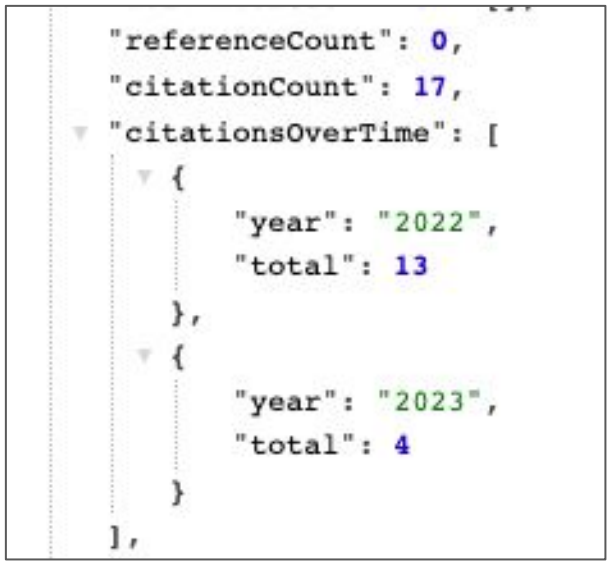
DataCite Commons: Repository



https://commons.datacite.org/repositories/nxrc8v

DataCite REST API

In the **“meta”** section of a DOI result:



from
https://api.datacite.org/dois/10.21227/781w-ef42

Part 4: The Global Data Citation Corpus

Data Citations: a role for everyone

DataCite

There is a role for all actors in the research process in supporting best practices for data citation

Researchers

- Deposit dataset at repositories that assign DOIs
- Include citations to the datasets they have used

Funders

- Encourage researchers to cite datasets they use - their own and others' - in their research outputs
- Consider information on datasets shared and their reuse as part of evaluation processes

Data repositories

- Collect citation information for datasets
- Include those citations in the metadata deposited with DataCite

Publishers

- Ensure data citations are included in the article, in a machine-readable format
- Index citations with Crossref

Data Citations: Completing the picture

Data citations through DOI metadata provide valuable information on data usage, but we know there are many more instances of data usage than we are currently capturing.

Researchers do not always cite the data they use

Some repositories (e.g. in life sciences & biomedicine) use accession numbers instead of DOIs

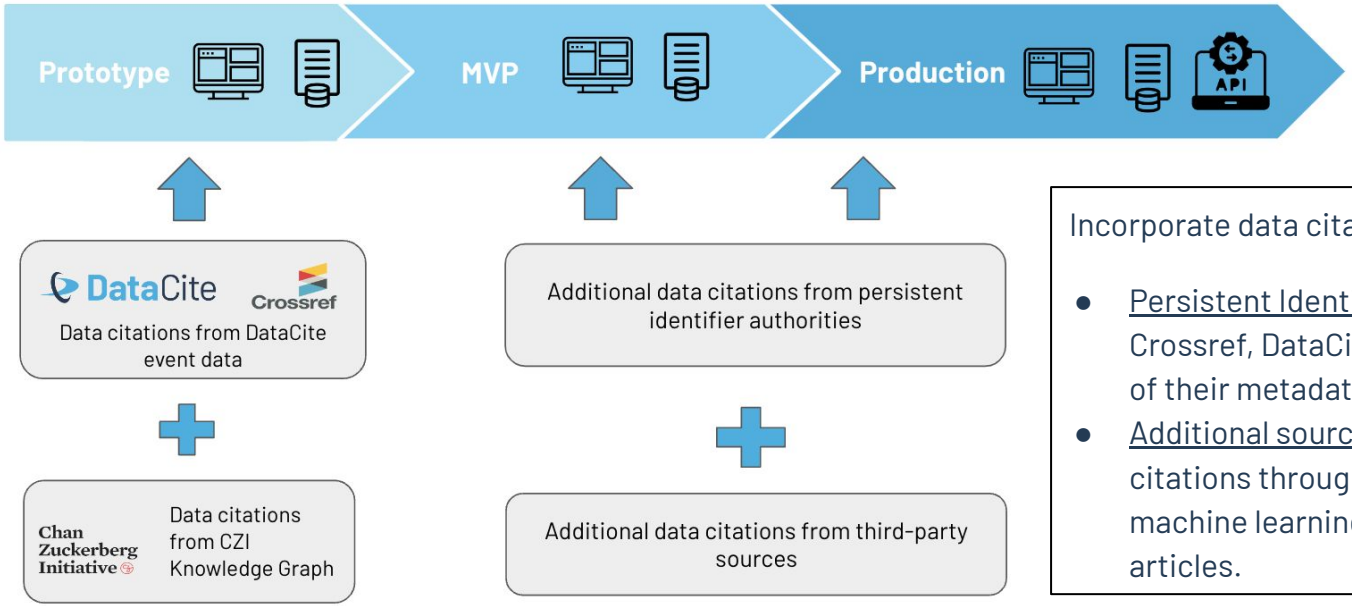
Citations and/or associated metadata may be lost when the publisher deposits in Crossref

Event data

In addition, some stakeholders have created approaches to find data citations that are not included in PID metadata, but this information exists in different platforms, often not publicly available

Open Global Data Citation Corpus

Goal: **Develop a comprehensive corpus that incorporates data citations from different sources into a centralized, publicly accessible community resource**



Incorporate data citations from diverse sources:

- Persistent Identifier (PID) authorities (e.g. Crossref, DataCite) that collect citations as part of their metadata deposit workflow.
- Additional sources that aggregate or discover citations through various techniques, such as machine learning and curation of full-text in articles.

Data Citation Corpus: Prototype

We are finalizing the prototype for the corpus, which will include data citations from DataCite event data and data mentions from CZI Science Knowledge Graph.

- Basic user interface to visualize the data with different filters
- Seed data available via a data dump

makedatacount.org/data-citation/
http://corpus.stage.datacite.org/dashboard

Takeaways

As a repository, you can be an active participant in providing citations (including data citations).

Ask depositors to include citations and follow up with subsequent citations when known.

Remember you can add citations in Fabrica, API at any time - including after the DOI is created

Watch our other trainings for more information on:

- DataCite Metadata (Part 1, Part 2)
- DataCite Fabrica
- DataCite Commons
- DataCite API

Resources

- DataCite Support Site
  - Contributing Citations and References
  - Consuming Citations and References
  - Views, downloads and citations in DataCite Commons
- DataCite Metadata Schema
- Make Data Count
  - Open Global Data Citation Corpus
  - Webinar recording: Building the Data Citation Corpus (February 2023)

DataCite

CONNECTING RESEARCH,
ADVANCING KNOWLEDGE

info@datacite.org

pidforum.org

datacite.org
blog.datacite.org

support.datacite.org
support@datacite.org

@datacite

DataCite

@datacite