

Harvesters Interest Group Meeting

28th of January 2026
Harvesters Interest Group Meeting



@datacite.org

(Re)-Introduction to the Harvesters Interest Group

Purpose of the Group

The Harvesters Interest Group is an opportunity to:

- learn about upcoming features from DataCite
- offer ideas for improvements and new features to DataCite staff
- collaborate with other harvesters and learn how they work
- communicate with DataCite Members contributing to the data file

Group Introductions

Mentimeter

Go to [menti.com](https://www.menti.com) and enter code **3440 3919**

Or use link: <https://www.menti.com/alx38bp9vwdc> (link in chat)

Today's Agenda

Agenda

1. Recent Release Highlights
2. Launching the DataCite Monthly Data File
3. Discussion topic: REST API Filters

Recent Release Highlights

Recent Release Highlights

- **DataCite's latest annual public data file is now available for 2025**, containing metadata for 108 million DataCite DOIs in Findable state that were registered up to the end of 2025.
 - Download the public data file via [our data files portal](#) and learn more in [our blog post](#).
- DataCite's public data file **will soon be available in the Dimensions BigQuery Lab** through a partnership with Digital Science.
 - Look out for an upcoming announcement on [our blog](#).

DataCite Monthly Data File

Launching the DataCite Monthly Data File Service



**DataCite REST
API**

**DataCite
OAI-PMH Service**

**DataCite Public
Data Files**

**DataCite Monthly
Data File**

Now Available

DataCite Monthly Data File

What's Included



The DataCite Monthly Data File includes:

- **A complete DataCite data file snapshot**, regenerated at the beginning of every month and available via AWS S3.
- **All publicly available DataCite DOI records** represented in JSON Lines files and separated by month updated.
- **Tabular reports** to identify records that are no longer publicly available and records that have moved between DataCite repositories.

Access to the Monthly Data File is available with DataCite Member, Consortium Organization, or Repository credentials.

DataCite Monthly Data File

Accessing the Monthly Data File

1. **Request temporary AWS credentials** from the DataCite REST API using DataCite Member, Consortium Organization, or Repository credentials.
2. Use AWS CLI, curl, AWS SDKs, and other access methods to **quickly and conveniently retrieve DataCite metadata** in bulk.

Read more about the service here:

<https://support.datacite.org/docs/datacite-monthly-data-file>

DataCite Monthly Data File

Accessing the Monthly Data File

1. Request temporary AWS credentials

REST API request to retrieve AWS credentials

```
curl --user datacite.blog:mock_password  
https://api.datacite.org/credentials/datafile?  
format=text
```

Response

```
[datacite-datafile]  
aws_access_key_id=EXAMPLE_KEY_ID  
aws_secret_access_key=EXAMPLE_SECRET_KEY  
aws_session_token=EXAMPLE_SESSION_TOKEN
```

2. Access the monthly data file

AWS CLI command to retrieve a list of all folders in the /dois/ directory

```
aws --profile datacite-datafile s3 ls  
s3://monthly-datafile.datacite.org/dois/
```

CLI output

```
PRE updated_2011-03/  
PRE updated_2011-04/  
...  
PRE updated_2025-11/  
PRE updated_2025-12/
```

DataCite Monthly Data File

Using the Monthly Data File



Download all DataCite DOI records in the file

```
aws --profile datacite-datafile s3 cp  
s3://monthly-datafile.datacite.org/does/ ./  
--recursive
```

Download all DataCite DOI records created or updated in December 2025

```
aws --profile datacite-datafile s3 cp  
s3://monthly-datafile.datacite.org/does/update  
d_2025-12/ ./ --recursive
```

Download a report of DataCite DOI records created or updated in December 2025

```
aws --profile datacite-datafile s3 cp  
s3://monthly-datafile.datacite.org/does/update  
d_2025-12/2025-12.csv.gz ./
```

Download the manifest file

```
aws --profile datacite-datafile s3 cp  
s3://monthly-datafile.datacite.org/MANIFEST ./
```

Learn more about access methods and how to build integrations:

<https://support.datacite.org/docs/datacite-monthly-data-file#other-access-methods-and-integration-support>

Options for Metadata Harvesting

Service	Availability	Update Frequency	Filters	Access pattern	Formats
REST API	All users, including unidentified users. Higher rate limits are available to identified and authenticated users.	Real-time	Yes	Request-based	JSON
OAI-PMH	All users, including unidentified users.	Real-time	Yes	Request-based	XML
Public Data File (Annual)	Identified users (must provide email address to access).	Annual	No	Bulk download	JSON Lines
Monthly Data File	Authenticated users.	Monthly	Records can be downloaded selectively by month updated	Bulk download	JSON Lines

REST API Filters

Available Filters

Filter Parameters

Parameter	Example values	Description
created	2025	Filter by the DOI date created (yyyy). Use commas to filter by multiple years (yyyy,yyyy).
registered	2025,2026	Filter by the DOI date registered (yyyy). Use commas to filter by multiple years (yyyy,yyyy).
provider-id	cern	Filter by a specific DataCite provider (Direct Member or Consortium Organization). Use commas to include multiple providers.
client-id	cern.zenodo	Filter by a specific DataCite client (repository). Use commas to include multiple repositories.
consortium-id	dcan	Filter by a specific DataCite consortium.
prefix	10.5061	Filter by a specific prefix. Use commas to include multiple prefixes.
schema-version	4	Filter by the specific schema version for the associated metadata.

Available Filters

Filter Parameters

Full parameter list in API Reference:

https://support.datacite.org/reference/get_dois

Parameter	Example values	Description
affiliation-id	https://ror.org/05dhe8b71	Search creators.affiliation.affiliationIdentifier and contributors.affiliation.affiliationIdentifier for a ROR ID.
funded-by	https://ror.org/01cwqze88	Search fundingReferences.funderIdentifier for a ROR ID. Results also include DOIs containing a Crossref Funder ID in fundingReferences.funderIdentifier corresponding to the ROR ID.
include-funder-child-organizations	true	When the `funded-by=` parameter is set to a ROR ID and `include-funder-child-organizations=true`, the returned list of DOIs will be filtered to DataCite resources funded by the organization identified by the ROR ID as well as all of its child organizations. Child organizations include direct descendants as well as children of children.
resource-type-id	image	Filter by the resourceTypeGeneral. Use commas to include multiple values.
resource-type	Map	Filter by the free text resourceType. Use commas to include multiple values.
affiliation	true	Set affiliation=true to see additional affiliation information such as the affiliation identifier.
publisher	true	Set publisher=true to see additional publisher information such as the publisher identifier.

Available Filters

"Query" Parameter

To search specific fields in the DOI metadata parameter, use the **query** parameter.

- All DataCite DOI metadata fields are available in JSON format.
- Nested fields are joined by a period (.).
- For example, to search for DOIs where the creator's identifier scheme is ORCID:

<https://api.datacite.org/doi?query=creators.nameIdentifiers.nameIdentifierScheme:ORCID>

Available fields listed in the support documentation:

<https://support.datacite.org/docs/api-queries#using-the-query-parameter>

Filtering Use Cases

- Retrieve DOIs by researchers affiliated with a given **institution**
- Retrieve DOIs funded by a give **funding organization**
- Retrieve DOIs of a specific **resource type**
- Retrieve DOIs registered/updated in a specific **date range**
- Retrieve DOI from a specific **repository**
- Retrieve DOIs from a given **country**

Example: Filter by Affiliation

At least one Creator or Contributor affiliated with the British Library:

- Using ROR ID (<https://ror.org/05dhe8b71>):
 - <https://api.datacite.org/does?affiliation=true&affiliation-id=https://ror.org/05dhe8b71>
 - 144 results
- Using ROR ID (<https://ror.org/05dhe8b71>) or affiliation string ("British Library"):
 - [https://api.datacite.org/does?affiliation=true&query=creators.affiliation.affiliationIdentifier:*05dhe8b71 OR contributors.affiliation.affiliationIdentifier:*05dhe8b71 OR creators.affiliation.name: "British Library " OR contributors.affiliation.name: "British Library "](https://api.datacite.org/does?affiliation=true&query=creators.affiliation.affiliationIdentifier:*05dhe8b71 OR contributors.affiliation.affiliationIdentifier:*05dhe8b71 OR creators.affiliation.name:)
 - 692 results
 - Affiliation strings are now case insensitive

Example: Filter by Funder

Contains a FundingReference for the National Institutes of Health (NIH):

- Using ROR ID (<https://ror.org/01cwqze88>):
 - <https://api.datacite.org/does?affiliation=true&funded-by=https://ror.org/01cwqze88>
 - 7,616 results
- Add [include-funder-child-organizations=true](#) to include child organizations of NIH:
 - <https://api.datacite.org/does?affiliation=true&funded-by=https://ror.org/01cwqze88&include-funder-child-organizations=true>
 - 59,617 results

More information about these new filter parameters:

<https://github.com/datacite/datacite-suggestions/discussions/221>

Discussion: Filtering Use Cases

Mentimeter

Go to [menti.com](https://www.menti.com) and enter code **3440 3919**

Or use link: <https://www.menti.com/alx38bp9vwdc> (link in chat)

Harvester Engagement Opportunities



- Join the DataCite Harvesters Interest Group [Google Group](#) (if you haven't already)
- Share how you reuse DataCite metadata for our [Metadata Harvesters](#) page:
- Contribute feedback on [DataCite Suggestions](#):
 - Respond to requests for feedback on new features and services
 - Share your ideas for how we can improve our services



CONNECTING RESEARCH,
ADVANCING KNOWLEDGE



info@datacite.org



datacite.org
datacite.org/blog



support.datacite.org
support@datacite.org



[@datacite](https://twitter.com/datacite)



[@datacite](https://facebook.com/datacite)



[@datacite](https://linkedin.com/company/datacite)



[DataCite](https://youtube.com/DataCite)