HELMHOLTZ PROJECT HERMES

HELMHOLTZ RICH METADATA SOFTWARE PUBLICATIONS

HERMES Kickoff Workshop 2021-11-12

ALFRED P. SLOAN FOUNDATION

The Dataverse Project - Supporting (research) code in dataset publications

HERMES (Helmholtz Rich Metadata Software Publication) kick-off workshop
November 12, 2021

Ana Trisovic, Harvard University
on behalf of the Dataverse Project team

- A free and open-source software platform to archive, share, and cite research data
  - Focus on data sharing and making data available
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community
  - 128 contributors to the software

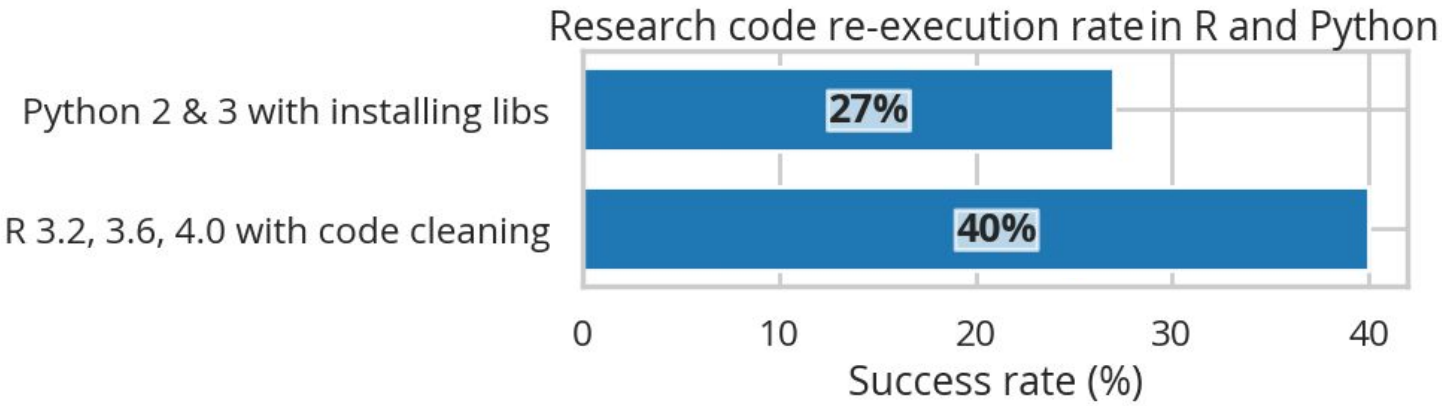
74 institutions around the globe run Dataverse installations as their official data repository

Data sharing at Dataverse

- Dataverse has data handling as its core strength
- Over the years, we see an increasing percentage of datasets with code
- Replication dataset - a bundle of data, code and other files needed to reproduce a published study
  - Journals like AJPS require data & code deposits in their collections

It is hard to re-execute published research code!

- Most code files fail when re-executed out-of-the-box, even with the pre-installation of used libraries [1,2].

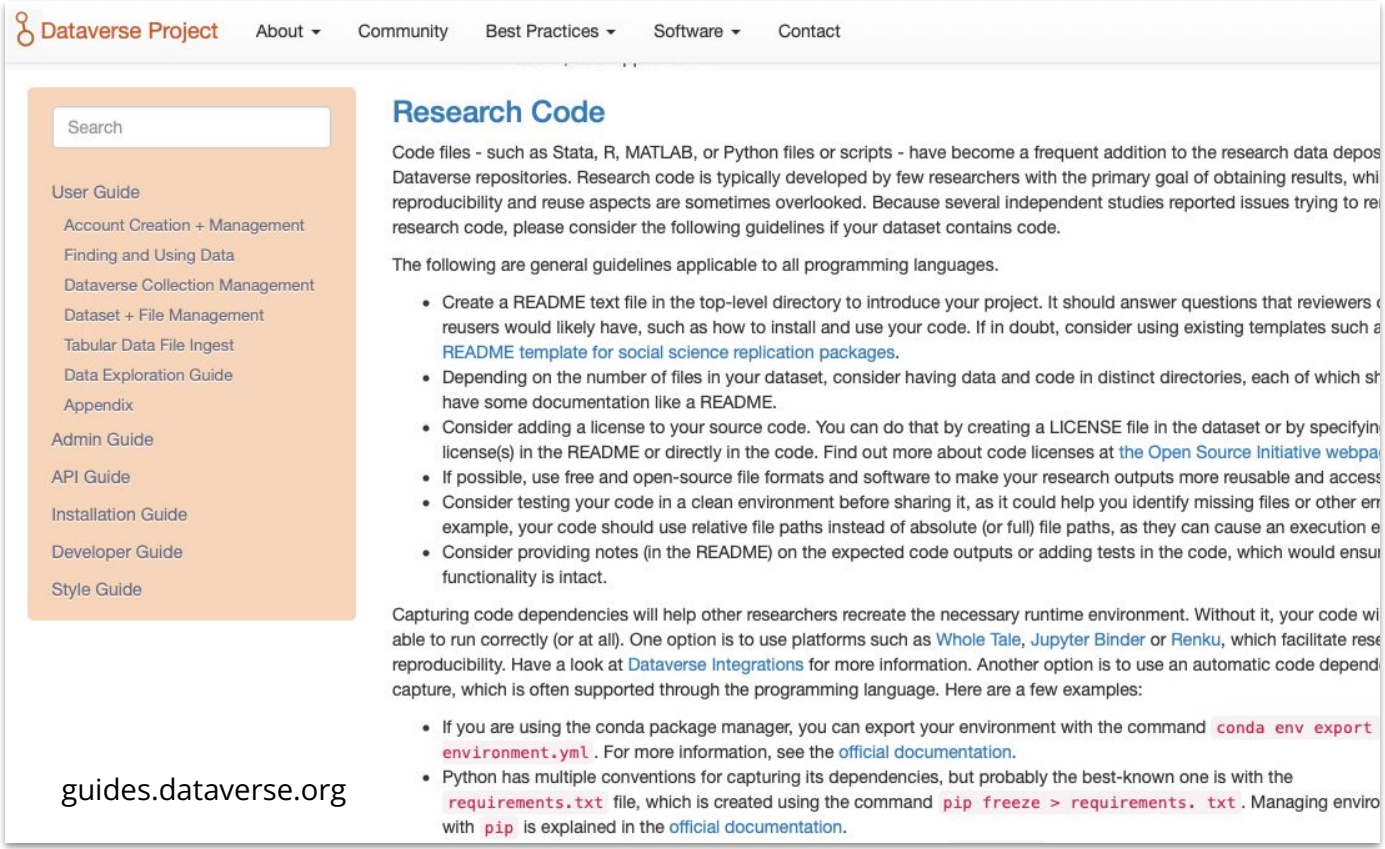


[1] Trisovic, Ana, et al. "Repository Approaches to Improving Quality of Shared Data and Code." Data 6.2 (2021): 15.
[2] Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

What can we do to support research code in Dataverse?

Docs: Official documentation and guidelines

The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.



Dataverse Project | About | Community | Best Practices | Software | Contact

Search

User Guide
- Account Creation + Management
- Finding and Using Data
- Dataverse Collection Management
- Dataset + File Management
- Tabular Data File Ingest
- Data Exploration Guide
- Appendix

Admin Guide

API Guide

Installation Guide

Developer Guide

Style Guide

Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to the research data depos
Dataverse repositories. Research code is typically developed by few researchers with the primary goal of obtaining results, whi
reproducibility and reuse aspects are sometimes overlooked. Because several independent studies reported issues trying to re
research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer questions that reviewers
reusers would likely have, such as how to install and use your code. If in doubt, consider using existing templates such a
README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct directories, each of which sh
have some documentation like a README.
- Consider adding a license to your source code. You can do that by creating a LICENSE file in the dataset or by specifyin
license(s) in the README or directly in the code. Find out more about code licenses at the Open Source Initiative webpa
- If possible, use free and open-source file formats and software to make your research outputs more reusable and access
- Consider testing your code in a clean environment before sharing it, as it could help you identify missing files or other er
example, your code should use relative file paths instead of absolute (or full) file paths, as they can cause an execution e
- Consider providing notes (in the README) on the expected code outputs or adding tests in the code, which would ensu
functionality is intact.

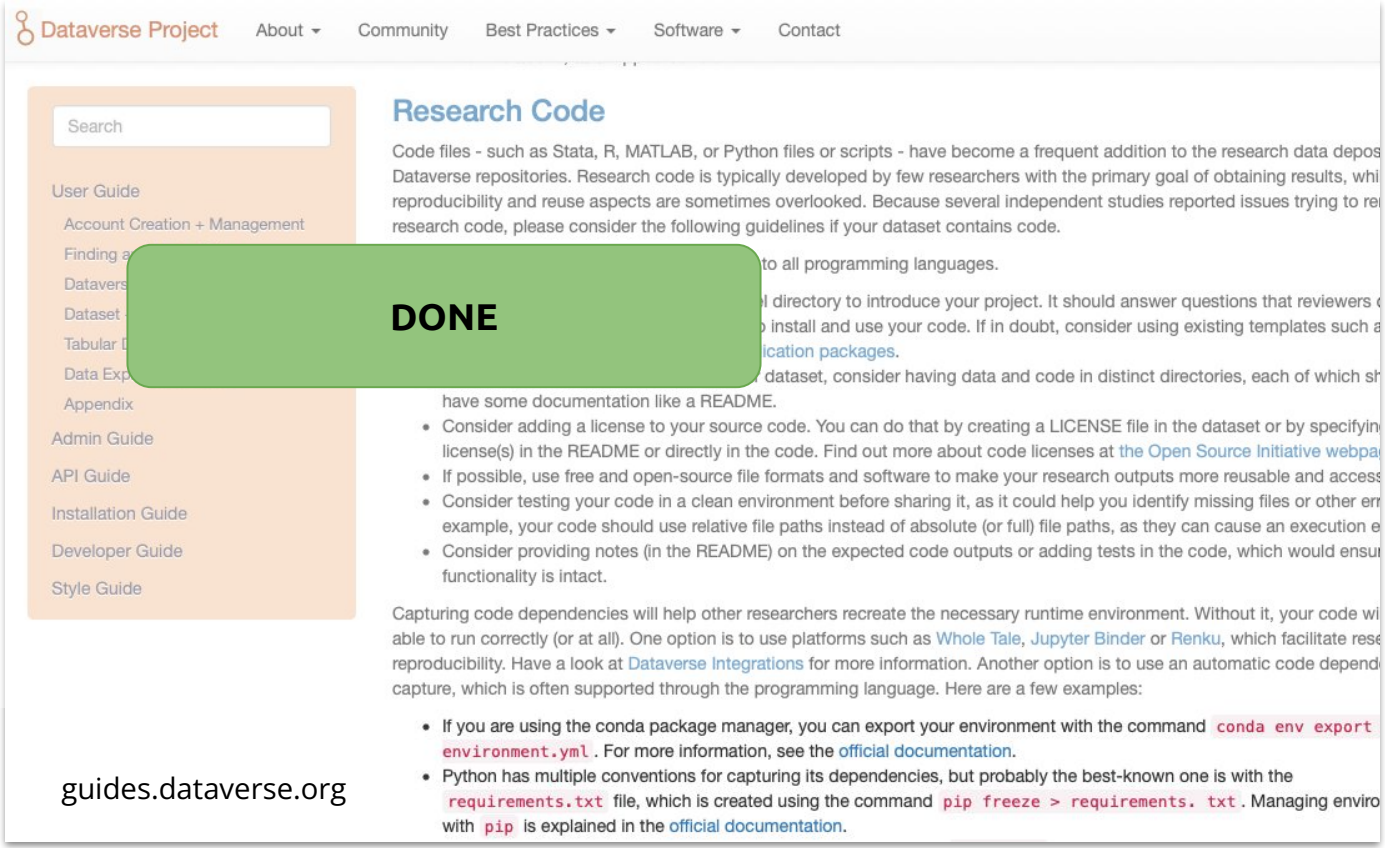
Capturing code dependencies will help other researchers recreate the necessary runtime environment. Without it, your code wi
able to run correctly (or at all). One option is to use platforms such as Whole Tale, Jupyter Binder or Renku, which facilitate rese
reproducibility. Have a look at Dataverse Integrations for more information. Another option is to use an automatic code depend
capture, which is often supported through the programming language. Here are a few examples:

- If you are using the conda package manager, you can export your environment with the command `conda env export environment.yml`. For more information, see the official documentation.
- Python has multiple conventions for capturing its dependencies, but probably the best-known one is with the `requirements.txt` file, which is created using the command `pip freeze > requirements. txt`. Managing enviro
with `pip` is explained in the official documentation.

guides.dataverse.org

Docs: Official documentation and guidelines

The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.

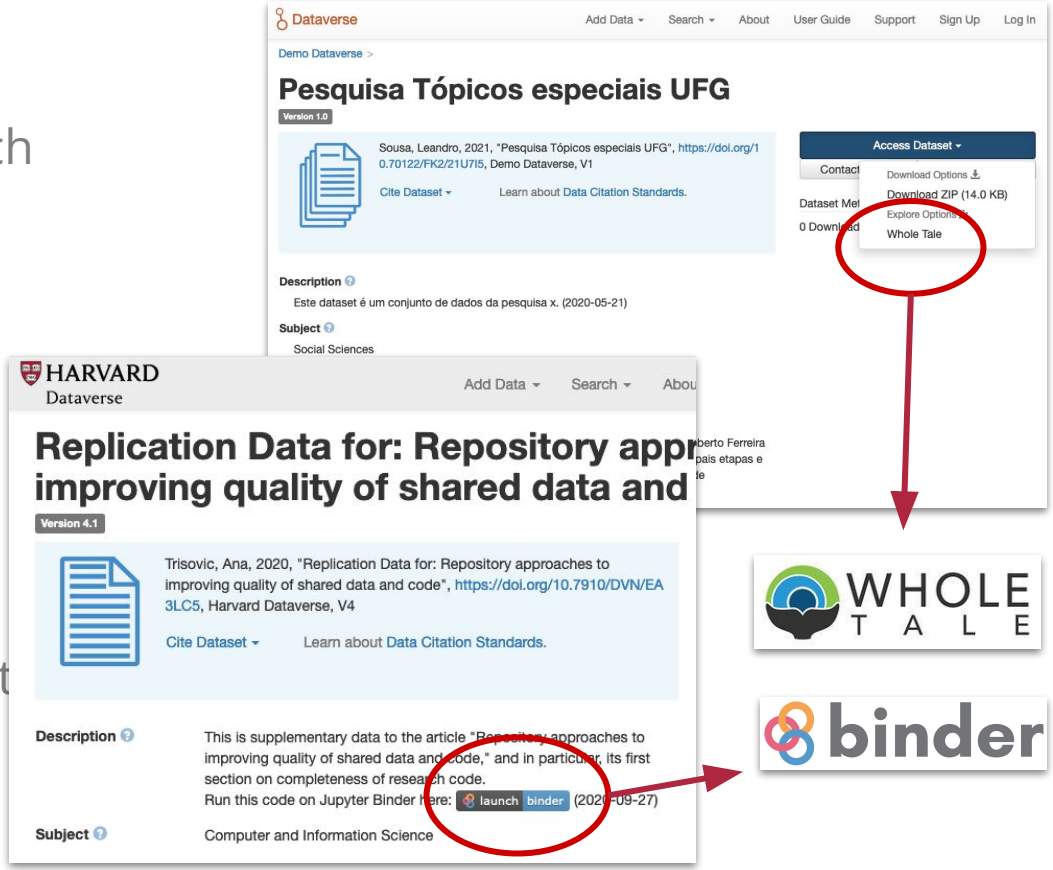


DONE

guides.dataverse.org

External tools: integration with cloud platforms

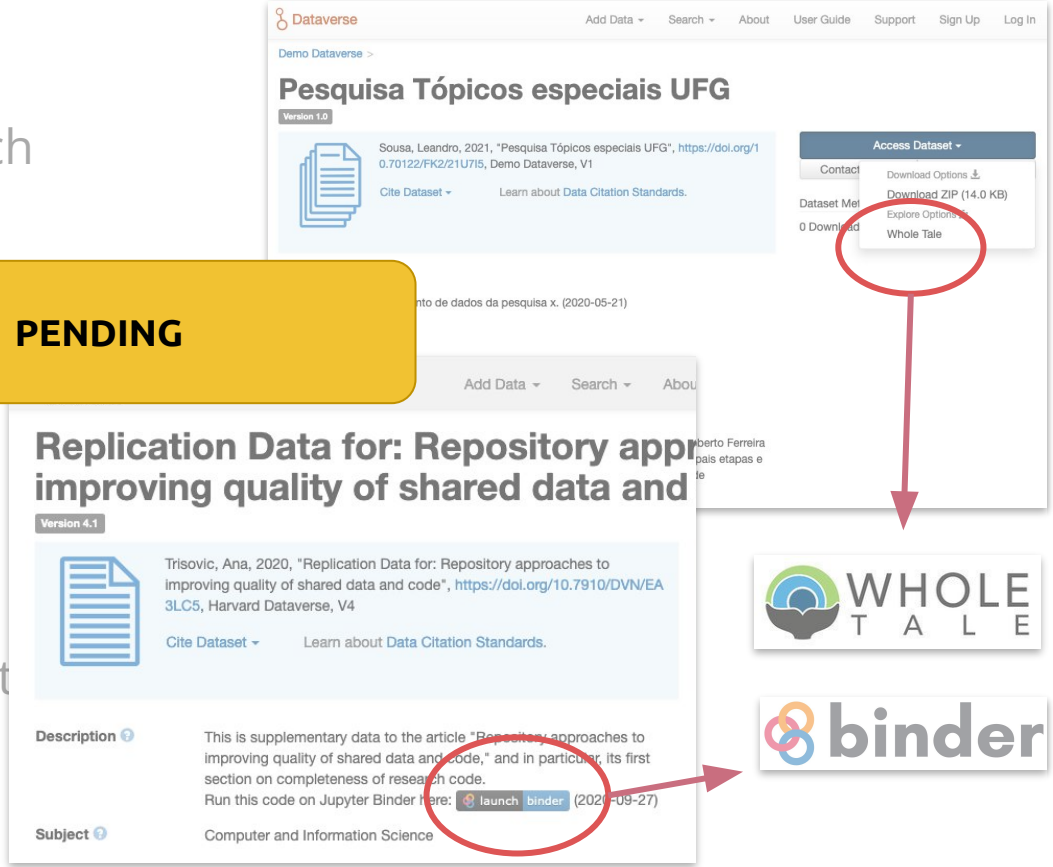
- New cloud platforms support collaborative work and research reproducibility by capturing necessary code dependencies from a web browser
- Integration with Dataverse:
  - Importing new research replication datasets
  - Exporting and reusing the existing replication dataset



External tools: integration with cloud platforms

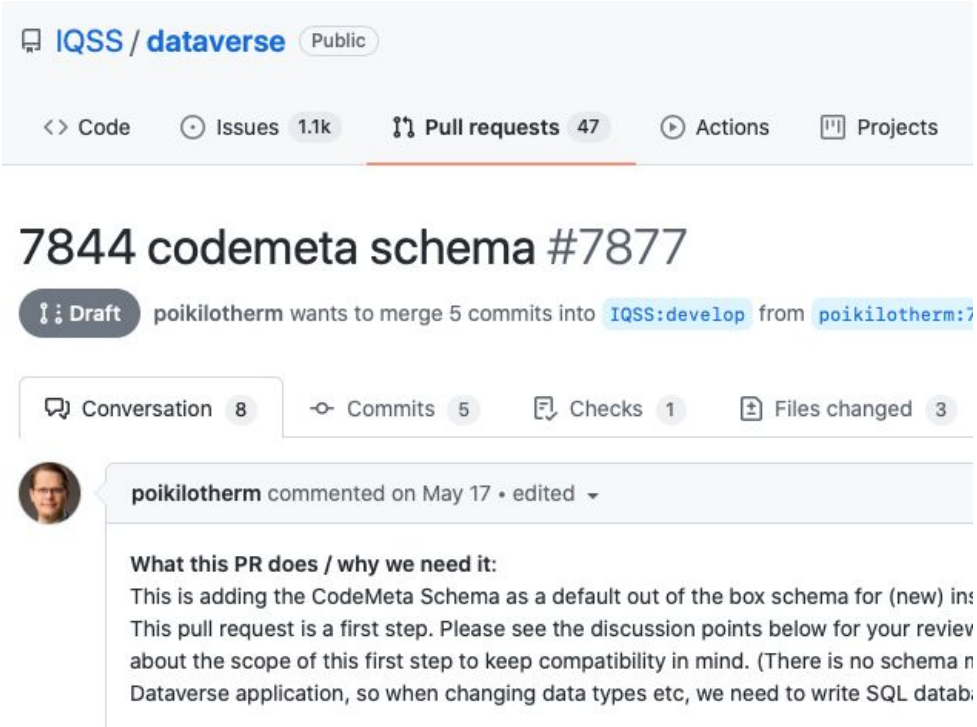
- New cloud platforms support collaborative work and research reproducibility by capturing necessary code depe
  from a web browser
- Integration with Dataverse:
  - Importing new research replication datasets
  - Exporting and reusing the existing replication dataset

PENDING



Metadata: Codemeta schema for code files

- Software metadata is necessary to document code files within the dataset.
  - Credit for academic software (citation)
  - Versions and dependencies of the software
  - Software licences



Metadata: Codemeta schema for code files

- Software metadata is necessary to document code files within the dataset.
  - Credit for acader software (citation)
  - Versions and dependencies of the software
  - Software licences

ONGOING

IQSS / dataverse Public

Code | Issues 1.1k | Pull requests 47 | Actions | Projects

hema #7877

rge 5 commits into IQSS:develop from poikilotherm:7

Conversation 8 | Commits 5 | Checks 1 | Files changed 3

poikilotherm commented on May 17 • edited

What this PR does / why we need it:

This is adding the CodeMeta Schema as a default out of the box schema for (new) ins
This pull request is a first step. Please see the discussion points below for your reviev
about the scope of this first step to keep compatibility in mind. (There is no schema n
Dataverse application, so when changing data types etc, we need to write SQL databa

Resource type: A new type for data & code

- When a dataset is published at Dataverse, its metadata is forwarded to DataCite, which facilitates its visibility on the web.
  - All deposits are `Dataset`s.
  - New resource type could be `Replication package`?

```
▼<title>
    Replication data for: Judging Under Public Pressure
  </title>
</titles>
<publisher>Harvard Dataverse</publisher>
<publicationYear>2021</publicationYear>
<resourceType resourceTypeGeneral="Dataset"/>
▼<descriptions>
  ▼<description descriptionType="Abstract">
    Replicating the tables in "Judging Under Public Pressure"
  </description>
</descriptions>
▼<contributors>
  ▼<contributor contributorType="ContactPerson">
```

DataCite
FIND, ACCESS, AND REUSE DATA

10.1 resourceTypeGeneral

Occurrence: 1
Definition: The general type of a resource.
Comment: **Values, examples, other constraints:**
Values:

- Audiovisual
- Book
- BookChapter
- Collection
- ComputationalNotebook
- ConferencePaper
- ConferenceProceeding
- DataPaper
- Dataset
- Dissertation
- Event
- Image
- InteractiveResource
- Journal
- JournalArticle
- Model
- OutputManagementPlan
- PeerReview
- PhysicalObject
- Preprint
- Report
- Service
- Software
- Sound
- Standard
- Text
- Workflow
- Other

https://support.datacite.org/docs/datacite-metadata-schema-v44-mandatory-properties#101-resourcetypegeneral

Resource type: A new type for data & code

- When a dataset is published at Dataverse, its metadata is forwarded to DataCite, which facilitates its visibility on the web.
  - All deposits are `Datasets`
  - New resource ty[illegible] `package`?

NEEDED

```
▼<title>
    Replication data for: Judging Under Public Pressure
  </title>
</titles>
<publisher>Harvard Dataverse</publisher>
<publicationYear>2021</publicationYear>
<resourceTypeGeneral="Dataset"/>
▼<descriptions>
  ▼<description descriptionType="Abstract">
    Replicating the tables in "Judging Under Public Pressure"
  </description>
</descriptions>
▼<contributors>
  ▼<contributor contributorType="ContactPerson">
```

DataCite
FIND, ACCESS, AND REUSE DATA

10.1 resourceTypeGeneral

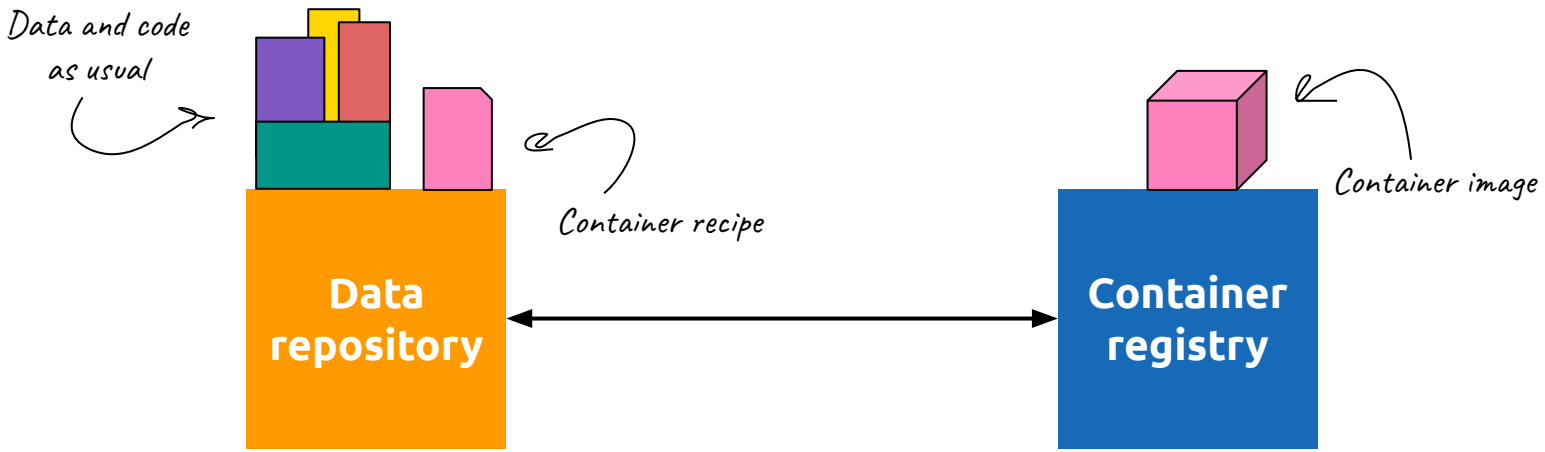
Occurrence: 1
Definition: The general type of a resource.
[illegible], examples, other constraints:
Values:

- Audiovisual
- Book
- BookChapter
- Collection
- ComputationalNotebook
- ConferencePaper
- ConferenceProceeding
- DataPaper
- Dataset
- Dissertation
- Event
- Image
- InteractiveResource
- Journal
- JournalArticle
- Model
- OutputManagementPlan
- PeerReview
- PhysicalObject
- Preprint
- Report
- Service
- Software
- Sound
- Standard
- Text
- Workflow
- Other

https://support.datacite.org/docs/datacite-metadata-schema-v44-mandatory-properties#101-resourcetypegeneral

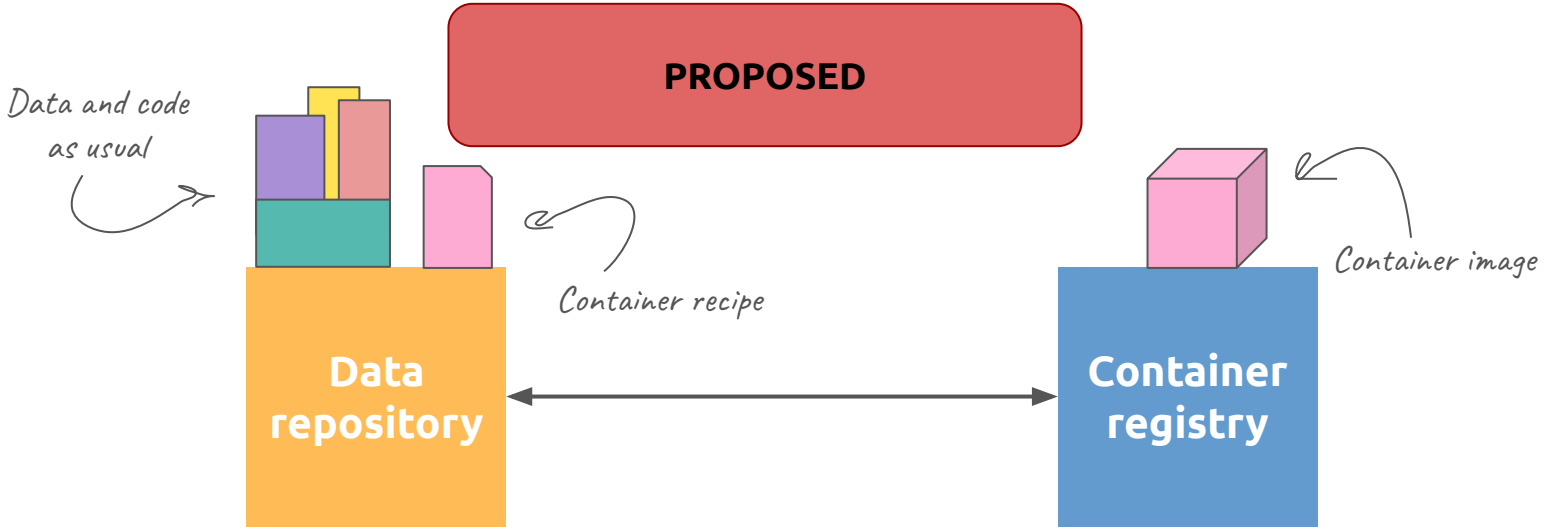
Additional: Support for virtual containers

- Dissemination of computational components, such as containers, with their metadata, tools and infrastructure
- Ongoing discussion at the Dataverse SWC (swc.gdcc.io) working group



Additional: Support for virtual containers

- Dissemination of computational components, such as containers, with their metadata, tools and infrastructure
- Ongoing discussion at the Dataverse SWC (swc.gdcc.io) working group



ALFRED P. SLOAN
FOUNDATION

Thank you!

Email: anatrisovic@g.harvard.edu
GitHub & Twitter: atrisovic
Dataverse Project: https://dataverse.org/contact

Thank you!

Where to learn more about project HERMES?

Stephan Druskat, DLR, PI, *@stdruskat*

Oliver Bertuch, FZJ, PI, *@poi_ki_lo_therm*

Guido Juckeland, HZDR, PI, *@GuidoJuckeland*

Oliver Knodel, HZDR, *@olikno1*

Tobias Schlauch, DLR, *@TobiasSchlauch*

- Find us on Twitter
- Write an email to team@software-metadata.pub
- Go to software-metadata.pub