

# HERMES Kickoff Workshop 2021-11-12

# 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

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) [ajps.org](#)

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

## Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Access Dataset -  
Contact Owner Share

Dataset Metrics  
1,092 Downloads

**Description** ⓘ  
How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position—even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

**Subject** ⓘ  
Social Sciences

**Keyword** ⓘ  
Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

**Related Publication** ⓘ  
Bisgaard, Martin, and Rune Slothuus. [date], "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. <http://ajps.org/>

**Notes** ⓘ  
This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the [Center for Open Science](#).



**Code, documentation and other files**

Files Metadata Terms Versions

Search this dataset...

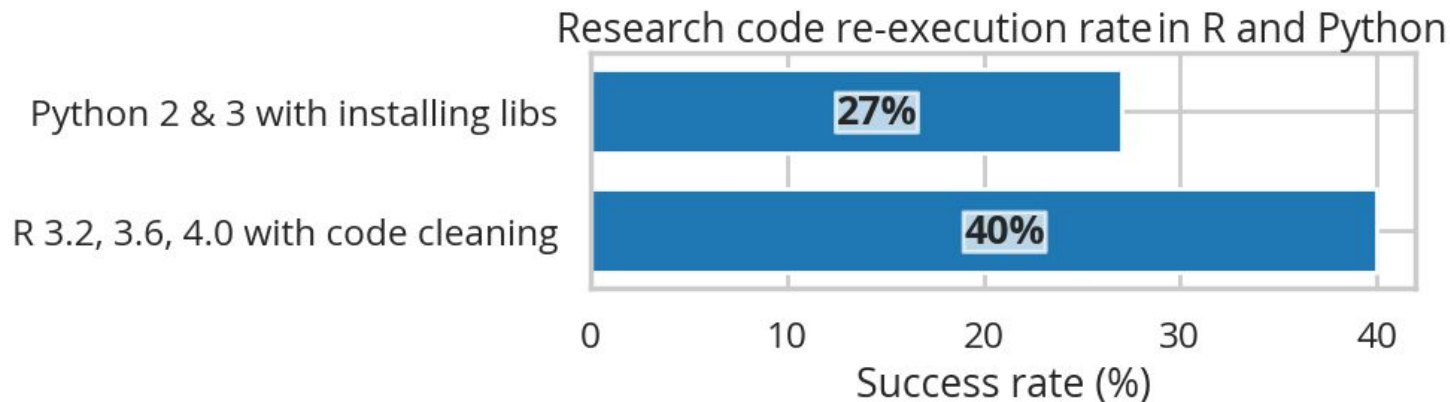
Filter by  
File Type: All Access: All

1 to 10 of 25 Files

<input type="checkbox"/>	<b>build_data.R</b> R Syntax - 12.1 KB Published Jun 29, 2020 56 Downloads MD5: a94...597	
<input type="checkbox"/>	<b>codebook ess.pdf</b> Adobe PDF - 508.8 KB Published Jun 29, 2020 46 Downloads	

# 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].



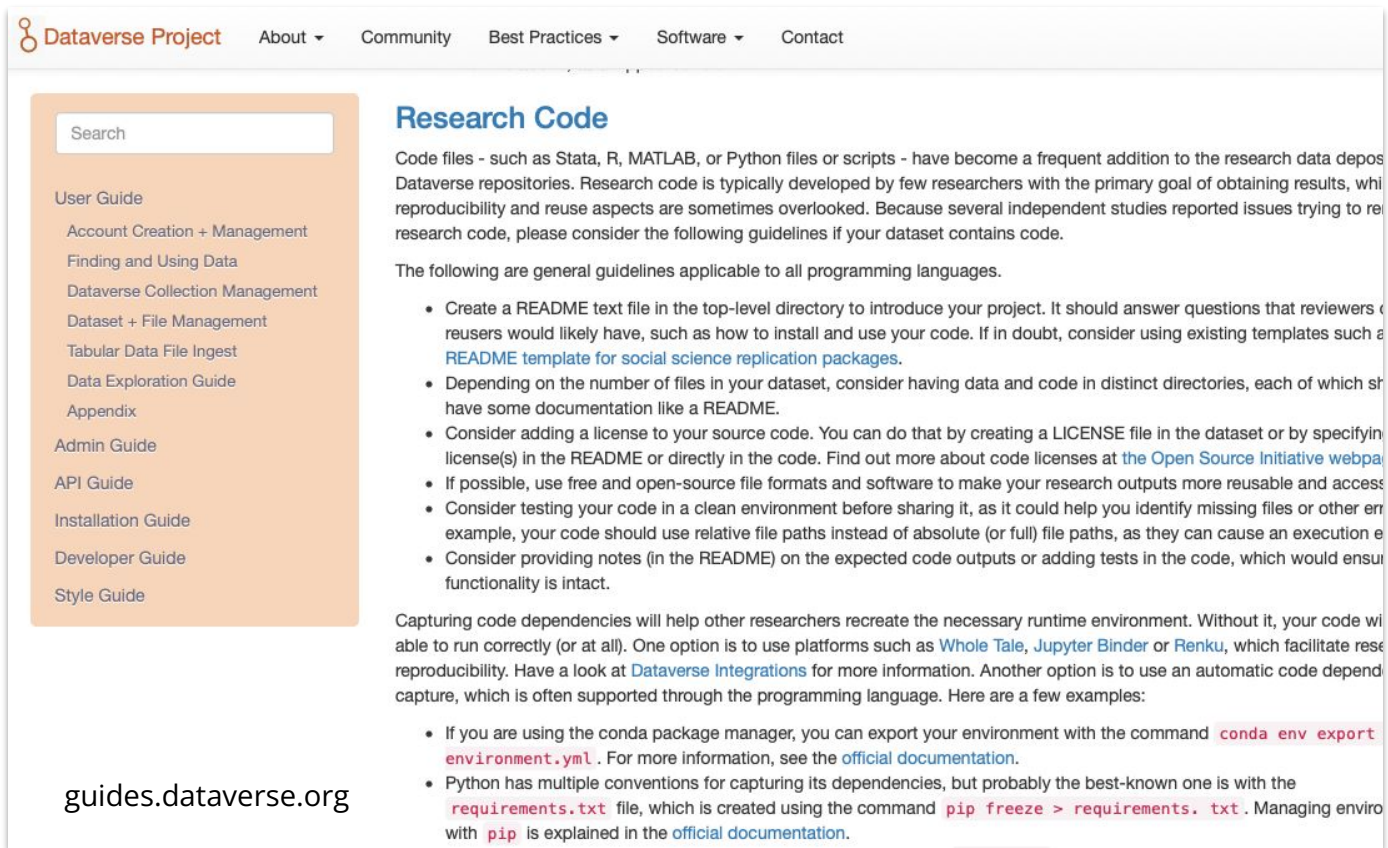
[1] Trisovic, Ana, et al. "Repository Approaches to Improving Quality of Shared Data and Code." Data 6.2 (2021): 15.

[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.



The screenshot shows the Dataverse Project website. The header includes the logo and navigation links: About, Community, Best Practices, Software, and Contact. A left sidebar contains a search bar and a list of links: 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, and Style Guide. The main content area features a 'Research Code' section with a paragraph explaining the importance of research code and a list of guidelines. Below this is a paragraph about capturing code dependencies and another list of examples.

**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 deposited in Dataverse repositories. Research code is typically developed by few researchers with the primary goal of obtaining results, while reproducibility and reuse aspects are sometimes overlooked. Because several independent studies reported issues trying to reuse 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 or reusers would likely have, such as how to install and use your code. If in doubt, consider using existing templates such as the [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 should 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 specifying license(s) in the README or directly in the code. Find out more about code licenses at the [Open Source Initiative website](#).
- If possible, use free and open-source file formats and software to make your research outputs more reusable and accessible.
- Consider testing your code in a clean environment before sharing it, as it could help you identify missing files or other errors. For example, your code should use relative file paths instead of absolute (or full) file paths, as they can cause an execution error.
- Consider providing notes (in the README) on the expected code outputs or adding tests in the code, which would ensure that the functionality is intact.

Capturing code dependencies will help other researchers recreate the necessary runtime environment. Without it, your code will not be able to run correctly (or at all). One option is to use platforms such as [Whole Tale](#), [Jupyter Binder](#) or [Renku](#), which facilitate research reproducibility. Have a look at [Dataverse Integrations](#) for more information. Another option is to use an automatic code dependency 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 --name <env_name> > 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 environments with `pip` is explained in the [official documentation](#).

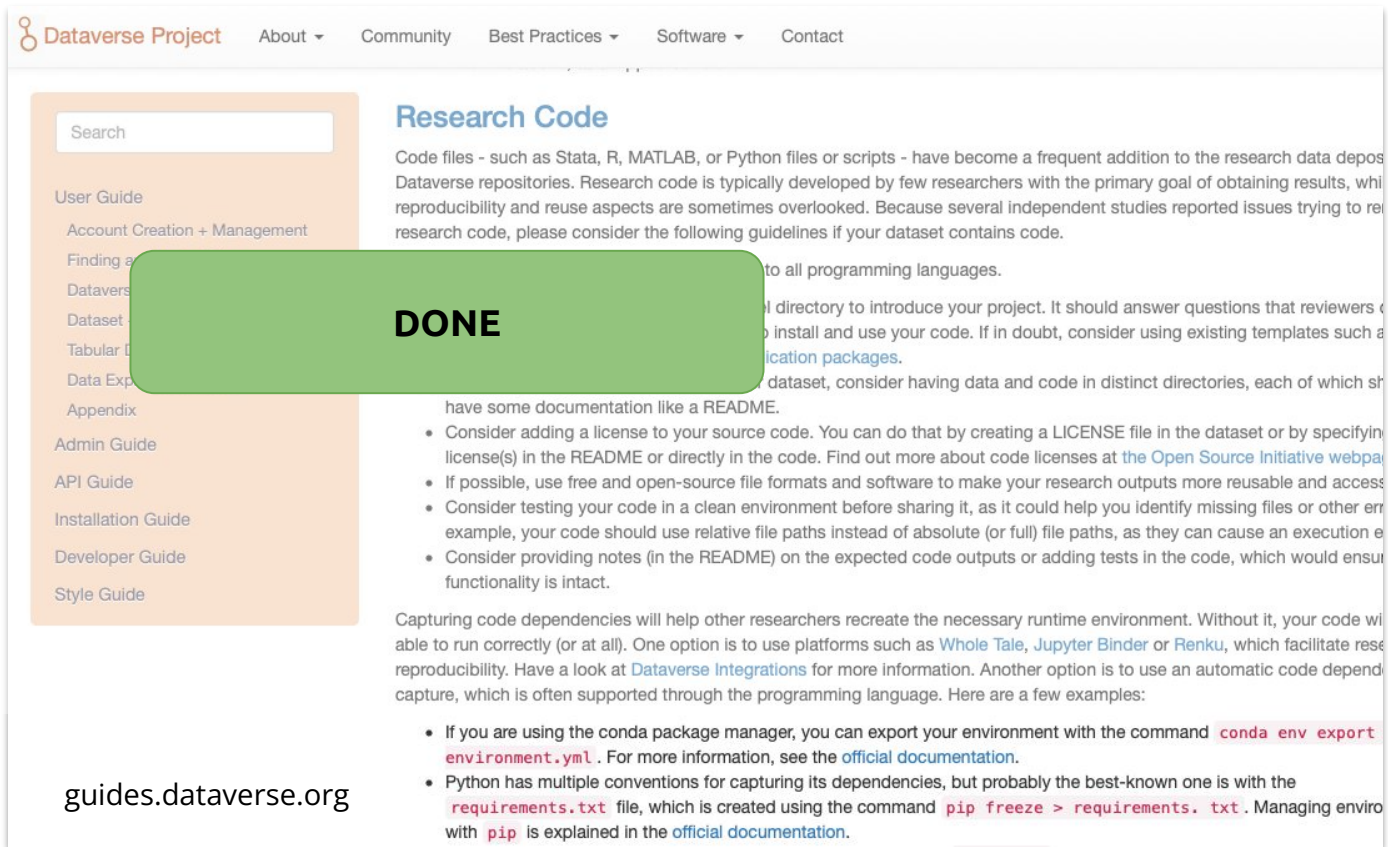
guides.dataverse.org



# Docs: Official documentation and guidelines

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

guides.dataverse.org



The screenshot shows the Dataverse Project website. The top navigation bar includes links for 'About', 'Community', 'Best Practices', 'Software', and 'Contact'. A left sidebar contains a search bar and a list of links: 'User Guide', 'Account Creation + Management', 'Finding a Dataset', 'Dataset', 'Tabular Data', 'Data Explorer', 'Appendix', 'Admin Guide', 'API Guide', 'Installation Guide', 'Developer Guide', and 'Style Guide'. A large green box with the word 'DONE' in white capital letters is overlaid on the sidebar. The main content area is titled 'Research Code' and contains text about code files and a list of guidelines for researchers. The guidelines include: considering adding a license, using free and open-source file formats, testing code in a clean environment, and providing notes on code outputs. The text also mentions capturing code dependencies and provides examples for using conda and pip.

**Research Code**

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

- Consider adding a license to your source code. You can do that by creating a LICENSE file in the dataset or by specifying the license(s) in the README or directly in the code. Find out more about code licenses at [the Open Source Initiative webpage](#).
- If possible, use free and open-source file formats and software to make your research outputs more reusable and accessible.
- Consider testing your code in a clean environment before sharing it, as it could help you identify missing files or other errors. For example, your code should use relative file paths instead of absolute (or full) file paths, as they can cause an execution error.
- Consider providing notes (in the README) on the expected code outputs or adding tests in the code, which would ensure that the functionality is intact.

Capturing code dependencies will help other researchers recreate the necessary runtime environment. Without it, your code will not be able to run correctly (or at all). One option is to use platforms such as [Whole Tale](#), [Jupyter Binder](#) or [Renku](#), which facilitate research reproducibility. Have a look at [Dataverse Integrations](#) for more information. Another option is to use an automatic code dependency 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 --name <env_name> --to <file>`. 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 environments with `pip` is explained in the [official documentation](#).

# 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

The diagram illustrates the workflow for integrating research datasets with cloud platforms. It features two screenshots of the Dataverse interface. The top screenshot shows a dataset titled "Pesquisa Tópicos especiais UFG" by Sousa, Leandro (2021). A red circle highlights the "Access Dataset" button, which opens a menu with options like "Download ZIP (14.0 KB)" and "Whole Tale". A red arrow points from this menu to the "WHOLE TALE" logo. The bottom screenshot shows a dataset titled "Replication Data for: Repository approaches to improving quality of shared data and code" by Trisovic, Ana (2020). A red circle highlights the "launch binder" button, with a red arrow pointing to the "binder" logo.

**Dataverse** Demo Dataverse >

### Pesquisa Tópicos especiais UFG

Version 1.0

Sousa, Leandro, 2021, "Pesquisa Tópicos especiais UFG", <https://doi.org/10.70122/FK2/21U715>, Demo Dataverse, V1

Cite Dataset - Learn about Data Citation Standards.

**Description** ⓘ  
Este dataset é um conjunto de dados da pesquisa x. (2020-05-21)

**Subject** ⓘ  
Social Sciences

**Access Dataset**

- Contact
- Download Options
- Download ZIP (14.0 KB)
- Explore Options
- Whole Tale

**HARVARD** Dataverse

### Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1

Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", <https://doi.org/10.7910/DVNEA3LCS>, Harvard Dataverse, V4

Cite Dataset - Learn about Data Citation Standards.

**Description** ⓘ  
This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.  
Run this code on Jupyter Binder here: [launch binder](#) (2020-09-27)

**Subject** ⓘ  
Computer and Information Science

**WHOLE TALE**

**binder**

# 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

**PENDING**

**Dataverse**

Demo Dataverse >

### Pesquisa Tópicos especiais UFG

Version 1.0

Sousa, Leandro, 2021, "Pesquisa Tópicos especiais UFG", <https://doi.org/10.70122/FK2/21U715>, Demo Dataverse, V1

Cite Dataset - Learn about Data Citation Standards.

Access Dataset -

- Contact
- Download Options
- Download ZIP (14.0 KB)
- Explore Options
- Whole Tale

Dataset Metadata Download

Conteúdo de dados da pesquisa x. (2020-05-21)

### Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1

Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", <https://doi.org/10.7910/DVNEA3LC5>, Harvard Dataverse, V4

Cite Dataset - Learn about Data Citation Standards.

Description ⓘ

This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.

Run this code on Jupyter Binder here: [launch binder](#) (2020-09-27)

Subject ⓘ

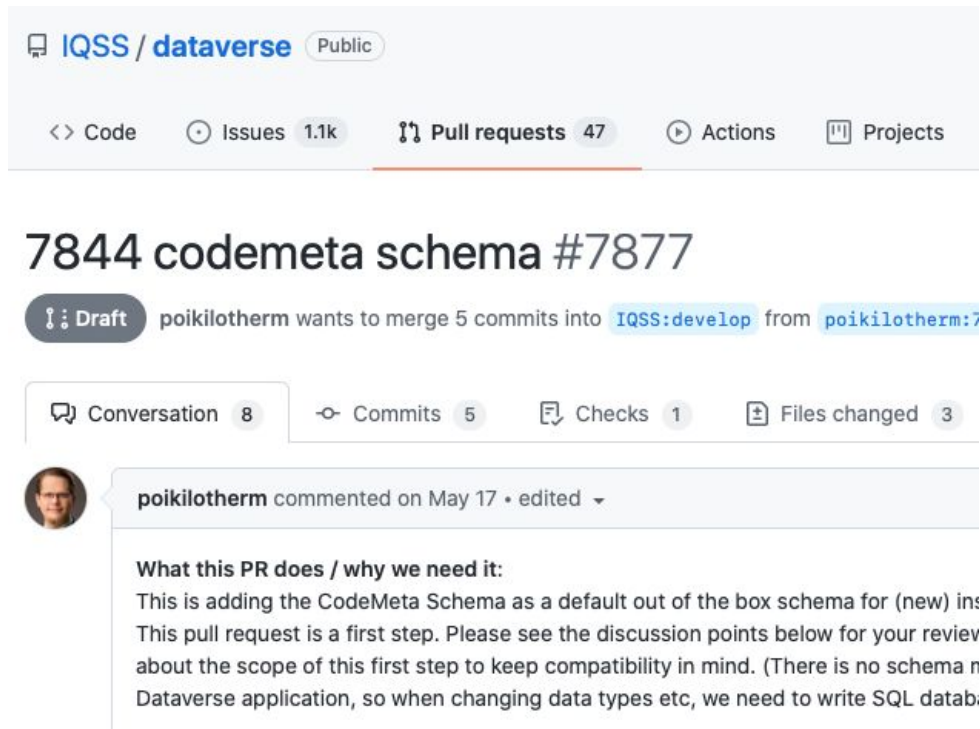
Computer and Information Science

**WHOLE TALE**

**binder**

# 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



The screenshot shows a GitHub interface for the `IQSS / dataverse` repository, which is marked as `Public`. The navigation bar includes links for `Code`, `Issues` (1.1k), `Pull requests` (47), `Actions`, and `Projects`. The main heading is `7844 codemeta schema #7877`, with a `Draft` label. The pull request description states: `poikilotherm wants to merge 5 commits into IQSS:develop from poikilotherm:7`. Below this, statistics show `8` conversations, `5` commits, `1` check, and `3` files changed. A comment by `poikilotherm` from May 17 is visible, containing the following text:

**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 review  
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 datab.

# 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

**ONGOING**

IQSS / dataverse Public

<> Code Issues 1.1k Pull requests 47 Actions Projects

Schema #7877

merge 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) installations. This pull request is a first step. Please see the discussion points below for your review about the scope of this first step to keep compatibility in mind. (There is no schema in the current Dataverse application, so when changing data types etc, we need to write SQL database updates.)



Occurrence: 1

Definition: The general type of a resource.

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

# 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 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" />
▼<description>
  ▼<description descriptionType="Abstract">
    Replicating the tables in "Judging Under Public Pressure"
  </description>
</descriptions>
▼<contributors>
  ▼<contributor contributorType="ContactPerson">

```



<https://support.datacite.org/docs/datacite-e-metadata-schema-v44-mandatory-properties#101-resourceTypeGeneral>

Occurrence: 1

Definition: The general type of a resource.

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

# 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 DataCite
  - New resource type package?

**NEEDED**

```

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

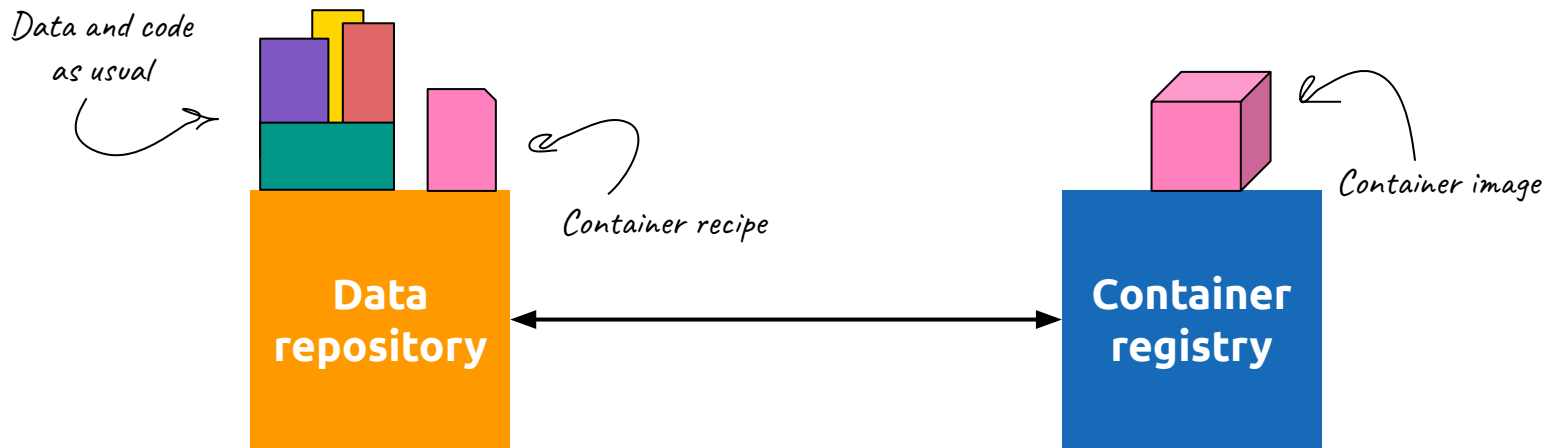
```



<https://support.datacite.org/docs/datacite-e-metadata-schema-v44-mandatory-properties#101-resource-type-general>

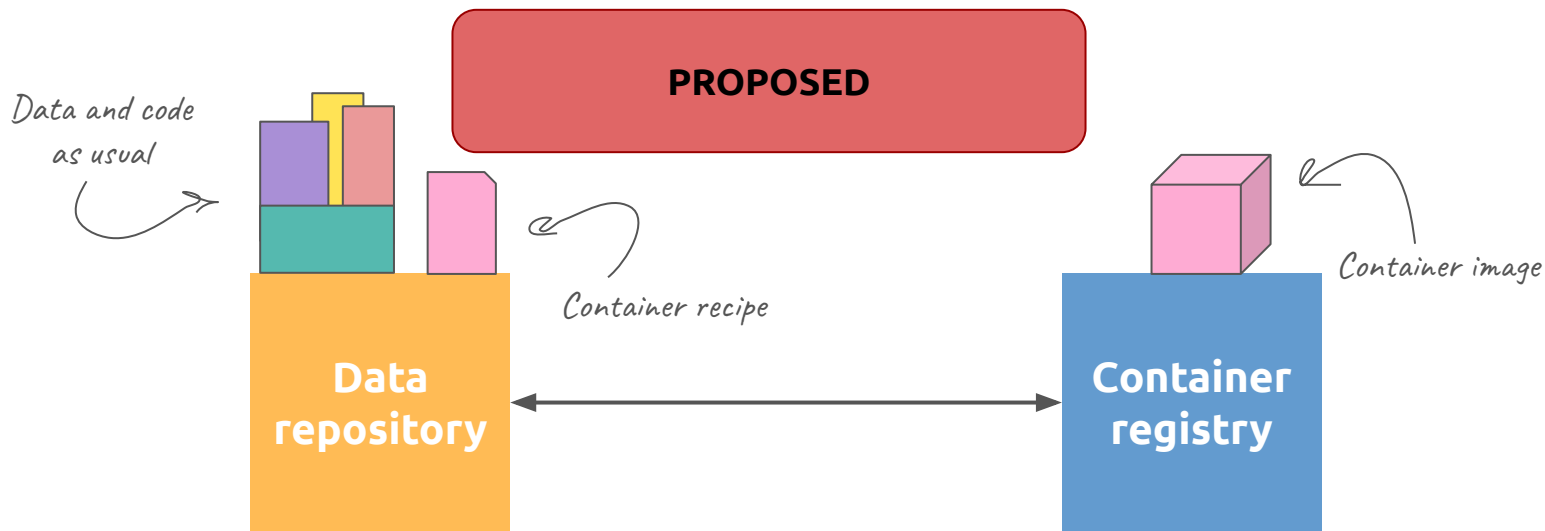
# Additional: Support for virtual containers

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



# Thank you!

Email: [anatrisovic@g.harvard.edu](mailto: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](mailto:team@software-metadata.pub)
- Go to [software-metadata.pub](http://software-metadata.pub)