# Datalad for containerized environments

*Release 1.2.5+8.g8ed8cca.dirty*

**DataLad team**

**Mar 22, 2024**

# CONTENTS

This extension equips DataLad's run/rerun functionality with the ability to transparently execute commands in containerized computational environments. On re-run, DataLad will automatically obtain any required container at the correct version prior execution.

CHAPTER

# ONE

# DOCUMENTATION

This is the technical documentation of the functionality and commands provided by this DataLad extension package. For an introduction to the general topic and a tutorial, please see the DataLad Handbook at https://handbook.datalad.org/r?containers.

- Documentation index
- *API reference*

## 1.1 Change log

```
 ____          _           _                 _
|  _ \   __ _ | |_   __ _ | |      __ _   __| |
| | | | / _` || __| / _` || |     / _` | / _` |
| |_| || (_| || |_ | (_| || |___ | (_| || (_| |
|____/  \__,_| \__| \__,_||_____| \__,_| \__,_|
                                      Container
```

This is a high level and scarce summary of the changes between releases. We would recommend to consult log of the DataLad git repository for more details.

### 1.1.1 1.1.2 (January 16, 2021) –

- Replace use of `mock` with `unittest.mock` as we do no longer support Python 2

### 1.1.2 1.1.1 (January 03, 2021) –

- Drop use of `Runner` (to be removed in datalad 0.14.0) in favor of `WitlessRunner`

### 1.1.3 1.1.0 (October 30, 2020) –

- Datalad version 0.13.0 or later is now required.
- In the upcoming 0.14.0 release of DataLad, the datalad special remote will have built-in support for "shub://" URLs. If `containers-add` detects support for this feature, it will now add the "shub://" URL as is rather than resolving the URL itself. This avoids registering short-lived URLs, allowing the image to be retrieved later with `datalad get`.
- `containers-run` learned to install necessary subdatasets when asked to execute a container from underneath an uninstalled subdataset.

### 1.1.4 1.0.1 (June 23, 2020) –

- Prefer `datalad.core.local.run` to `datalad.interface.run`. The latter has been marked as obsolete since DataLad v0.12 (our minimum requirement) and will be removed in DataLad's next feature release.

### 1.1.5 1.0.0 (Feb 23, 2020) – not-as-a-shy-one

Extension is pretty stable so releasing as 1. MAJOR release, so we could start tracking API breakages and enhancements properly.

- Drops support for Python 2 and DataLad prior 0.12

### 1.1.6 0.5.2 (Nov 12, 2019) –

#### Fixes

- The Docker adapter unconditionally called `docker run` with `--interactive` and `--tty` even when stdin was not attached to a TTY, leading to an error.

### 1.1.7 0.5.1 (Nov 08, 2019) –

#### Fixes

- The Docker adapter, which is used for the "dhub://" URL scheme, assumed the Python executable was spelled "python".
- A call to DataLad's `resolve_path` helper assumed a string return value, which isn't true as of the latest DataLad release candidate, 0.12.0rc6.

### 1.1.8 0.5.0 (Jul 12, 2019) – damn-you-malicious-users

#### New features

- The default result renderer for `containers-list` is now a custom renderer that includes the container name in the output.

### Fixes

- Temporarily skip two tests relying on SingularityHub – it is down.

## 1.1.9 0.4.0 (May 29, 2019) – run-baby-run

The minimum required DataLad version is now 0.11.5.

### New features

- The call format gained the "{img_dspath}" placeholder, which expands to the relative path of the dataset that contains the image. This is useful for pointing to a wrapper script that is bundled in the same subdataset as a container.
- `containers-run` now passes the container image to `run` via its `extra_inputs` argument so that a run command's "{inputs}" field is restricted to inputs that the caller explicitly specified.
- During execution, `containers-run` now sets the environment variable `DATALAD_CONTAINER_NAME` to the name of the container.

### Fixes

- `containers-run` mishandled paths when called from a subdirectory.
- `containers-run` didn't provide an informative error message when `cmdexec` contained an unknown placeholder.
- `containers-add` ignores the `--update` flag when the container doesn't yet exist, but it confusingly still used the word "update" in the commit message.

## 1.1.10 0.3.1 (Mar 05, 2019) – Upgrayeddd

### Fixes

- `containers-list` recursion actually does recursion.

## 1.1.11 0.3.0 (Mar 05, 2019) – Upgrayedd

### API changes

- `containers-list` no longer lists containers from subdatasets by default. Specify `--recursive` to do so.
- `containers-run` no longer considers subdataset containers in its automatic selection of a container name when no name is specified. If the current dataset has one container, that container is selected. Subdataset containers must always be explicitly specified.

### New features

- `containers-add` learned to update a previous container when passed `--update`.
- `containers-add` now supports Singularity's "docker://" scheme in the URL.
- To avoid unnecessary recursion into subdatasets, `containers-run` now decides to look for containers in subdatasets based on whether the name has a slash (which is true of all subdataset containers).

## 1.1.12 0.2.2 (Dec 19, 2018) – The more the merrier

- list/use containers recursively from installed subdatasets
- Allow to specify container by path rather than just by name
- Adding a container from local filesystem will copy it now

## 1.1.13 0.2.1 (Jul 14, 2018) – Explicit lyrics

- Add support `datalad run --explicit`.

## 1.1.14 0.2 (Jun 08, 2018) – Docker

- Initial support for adding and running Docker containers.
- Add support `datalad run --sidecar`.
- Simplify storage of `call_fmt` arguments in the Git config, by benefiting from `datalad run` being able to work with single-string compound commands.

## 1.1.15 0.1.2 (May 28, 2018) – The docs

- Basic beginner documentation

## 1.1.16 0.1.1 (May 22, 2018) – The fixes

### New features

- Add container images straight from singularity-hub, no need to manually specify `--call-fmt` arguments.

### API changes

- Use "name" instead of "label" for referring to a container (e.g. `containers-run -n ...` instead of `containers-run -l`.

#### Fixes

- Pass relative container path to `datalad run`.
- `containers-run` no longer hides `datalad run` failures.

### 1.1.17 0.1 (May 19, 2018) – The Release

- Initial release with basic functionality to add, remove, and list containers in a dataset, plus a `run` command wrapper that injects the container image as an input dependency of a command call.

## 1.2 Acknowledgments

DataLad development is being performed as part of a US-German collaboration in computational neuroscience (CRCNS) project "DataGit: converging catalogues, warehouses, and deployment logistics into a federated 'data distribution'" (Halchenko/Hanke), co-funded by the US National Science Foundation (NSF 1429999) and the German Federal Ministry of Education and Research (BMBF 01GQ1411). Additional support is provided by the German federal state of Saxony-Anhalt and the European Regional Development Fund (ERDF), Project: Center for Behavioral Brain Sciences, Imaging Platform

DataLad is built atop the git-annex software that is being developed and maintained by Joey Hess.

## 1.3 Metadata Extraction

If datalad-metalad extension is installed, *datalad-container* can extract metadata from singularity containers images.

(It is recommended to use a tool like *jq* if you would like to read the output yourself.)

### 1.3.1 Singularity Inspect

Adds metadata gathered from *singularity inspect* and the version of *singularity* or *apptainer*.

For example:

(From the ReproNim/containers repository)

*datalad meta-extract -d . container_inspect images/bids/bids-pymvpa--1.0.2.sing | jq*

```
{
  "type": "file",
  "dataset_id": "b02e63c2-62c1-11e9-82b0-52540040489c",
  "dataset_version": "9ed0a39406e518f0309bb665a99b64dec719fb08",
  "path": "images/bids/bids-pymvpa--1.0.2.sing",
  "extractor_name": "container_inspect",
  "extractor_version": "0.0.1",
  "extraction_parameter": {},
  "extraction_time": 1680097317.7093463,
  "agent_name": "Austin Macdonald",
  "agent_email": "austin@dartmouth.edu",
  "extracted_metadata": {
    "@id": "datalad:SHA1-s993116191--cc7ac6e6a31e9ac131035a88f699dfcca785b844",
```

(continues on next page)

(continued from previous page)

```
    "type": "file",
    "path": "images/bids/bids-pymvpa--1.0.2.sing",
    "content_byte_size": 0,
    "comment": "SingularityInspect extractor executed at 1680097317.6012993",
    "container_system": "apptainer",
    "container_system_version": "1.1.6-1.fc37",
    "container_inspect": {
      "data": {
        "attributes": {
          "labels": {
            "org.label-schema.build-date": "Thu,_19_Dec_2019_14:58:41_+0000",
            "org.label-schema.build-size": "2442MB",
            "org.label-schema.schema-version": "1.0",
            "org.label-schema.usage.singularity.deffile": "Singularity.bids-pymvpa--1.0.2
↪",
            "org.label-schema.usage.singularity.deffile.bootstrap": "docker",
            "org.label-schema.usage.singularity.deffile.from": "bids/pymvpa:v1.0.2",
            "org.label-schema.usage.singularity.version": "2.5.2-feature-squashbuild-
↪secbuild-2.5.6e68f9725"
          }
        }
      },
      "type": "container"
    }
  }
}
```

CHAPTER

# TWO

# API REFERENCE

## 2.1 Command manuals

### 2.1.1 datalad containers-add

#### Synopsis

```
datalad containers-add [-h] [-u URL] [-d DATASET] [--call-fmt FORMAT] [-i IMAGE] [--
↪update] [--extra-input FILE] [--version] NAME
```

#### Description

Add a container to a dataset

#### Options

#### NAME

The name to register the container under. This also determines the default location of the container image within the dataset. Constraints: value must be a string

#### -h, -\-help, -\-help-np

show this help message. --help-np forcefully disables the use of a pager for displaying the help message

#### -u URL, -\-url URL

A URL (or local path) to get the container image from. If the URL scheme is one recognized by Singularity (e.g., 'shub://neurodebian/dcm2niix:latest' or 'docker://debian:stable-slim'), a command format string for Singularity-based execution will be auto-configured when --call-fmt is not specified. For Docker- based container execution with the URL scheme 'dhub://', the rest of the URL will be interpreted as the argument to 'docker pull', the image will be saved to a location specified by NAME, and the call format will be auto-configured to run docker, unless overwritten. The auto-configured call to docker run mounts the CWD to '/tmp' and sets the working directory to '/tmp'. Constraints: value must be a string or value must be NONE

### -d *DATASET*, -\-dataset *DATASET*

specify the dataset to add the container to. If no dataset is given, an attempt is made to identify the dataset based on the current working directory. Constraints: Value must be a Dataset or a valid identifier of a Dataset (e.g. a path) or value must be NONE

### -\-call-fmt FORMAT

Command format string indicating how to execute a command in this container, e.g. "singularity exec {img} {cmd}". Where '{img}' is a placeholder for the path to the container image and '{cmd}' is replaced with the desired command. Additional placeholders: '{img_dspath}' is relative path to the dataset containing the image, '{img_dirpath}' is the directory containing the '{img}'. '{python}' expands to the path of the Python executable that is running the respective DataLad session, for example a 'datalad containers-run' command. Constraints: value must be a string or value must be NONE

### -i IMAGE, -\-image IMAGE

Relative path of the container image within the dataset. If not given, a default location will be determined using the NAME argument. Constraints: value must be a string or value must be NONE

### -\-update

Update the existing container for NAME. If no other options are specified, URL will be set to 'updateurl', if configured. If a container with *name* does not already exist, this option is ignored.

### -\-extra-input FILE

Additional file the container invocation depends on (e.g. overlays used in --call-fmt). Can be specified multiple times. Similar to --call-fmt, the placeholders {img_dspath} and {img_dirpath} are available. Will be stored in the dataset config and later added alongside the container image to the EXTRA_INPUTS field in the run-record and thus automatically be fetched when needed.

### -\-version

show the module and its version which provides the command

### Authors

datalad is developed by The DataLad Team and Contributors <team@datalad.org>.

## 2.1.2 datalad containers-remove

### Synopsis

```
datalad containers-remove [-h] [-d DATASET] [-i] [--version] NAME
```

### Description

Remove a known container from a dataset

This command is only removing a container from the committed Dataset configuration (configuration scope `branch`). It will not modify any other configuration scopes.

This command is *not* dropping the container image associated with the removed record, because it may still be needed for other dataset versions. In order to drop the container image, use the 'drop' command prior to removing the container configuration.

### Options

#### NAME

name of the container to remove. Constraints: value must be a string

#### -h, -\-help, -\-help-np

show this help message. --help-np forcefully disables the use of a pager for displaying the help message

#### -d *DATASET*, -\-dataset *DATASET*

specify the dataset from removing a container. If no dataset is given, an attempt is made to identify the dataset based on the current working directory. Constraints: Value must be a Dataset or a valid identifier of a Dataset (e.g. a path) or value must be NONE

#### -i, -\-remove-image

if set, remove container image as well. Even with this flag, the container image content will not be dropped. Use the 'drop' command explicitly before removing the container configuration.

#### -\-version

show the module and its version which provides the command

### Authors

datalad is developed by The DataLad Team and Contributors <team@datalad.org>.

## 2.1.3 datalad containers-list

### Synopsis

```
datalad containers-list [-h] [-d DATASET] [-r] [--contains PATH] [--version]
```

### Description

List containers known to a dataset

### Options

#### -h, -\-help, -\-help-np

show this help message. --help-np forcefully disables the use of a pager for displaying the help message

#### -d *DATASET*, -\-dataset *DATASET*

specify the dataset to query. If no dataset is given, an attempt is made to identify the dataset based on the current working directory. Constraints: Value must be a Dataset or a valid identifier of a Dataset (e.g. a path) or value must be NONE

#### -r, -\-recursive

if set, recurse into potential subdatasets.

#### -\-contains PATH

when operating recursively, restrict the reported containers to those from subdatasets that contain the given path (i.e. the subdatasets that are reported by datalad subdatasets --contains=PATH). Top-level containers are always reported.

#### -\-version

show the module and its version which provides the command

### Authors

datalad is developed by The DataLad Team and Contributors <team@datalad.org>.

## 2.1.4 datalad containers-run

### Synopsis

```
datalad containers-run [-h] [-n NAME] [-d DATASET] [-i PATH] [-o PATH] [-m MESSAGE] [--
↪expand {inputs|outputs|both}] [--explicit] [--sidecar {yes|no}] [--version] ...
```

### Description

Drop-in replacement of 'run' to perform containerized command execution

Container(s) need to be configured beforehand (see containers-add). If no container is specified and only one container is configured in the current dataset, it will be selected automatically. If more than one container is registered in the current dataset or to access containers from subdatasets, the container has to be specified.

A command is generated based on the input arguments such that the container image itself will be recorded as an input dependency of the command execution in the RUN record in the git history.

During execution the environment variable DATALAD_CONTAINER_NAME is set to the name of the used container.

### Options

#### COMMAND

command for execution. A leading '--' can be used to disambiguate this command from the preceding options to DataLad.

#### -h, -\-help, -\-help-np

show this help message. --help-np forcefully disables the use of a pager for displaying the help message

#### -n NAME, -\-container-name NAME

Specify the name of or a path to a known container to use for execution, in case multiple containers are configured.

#### -d *DATASET*, -\-dataset *DATASET*

specify the dataset to record the command results in. An attempt is made to identify the dataset based on the current working directory. If a dataset is given, the command will be executed in the root directory of this dataset. Constraints: Value must be a Dataset or a valid identifier of a Dataset (e.g. a path) or value must be NONE

### -i PATH, -\-input PATH

A dependency for the run. Before running the command, the content for this relative path will be retrieved. A value of "." means "run datalad get .". The value can also be a glob. This option can be given more than once.

### -o PATH, -\-output PATH

Prepare this relative path to be an output file of the command. A value of "." means "run datalad unlock ." (and will fail if some content isn't present). For any other value, if the content of this file is present, unlock the file. Otherwise, remove it. The value can also be a glob. This option can be given more than once.

### -m MESSAGE, -\-message MESSAGE

a description of the state or the changes made to a dataset. Constraints: value must be a string or value must be NONE

### -\-expand {inputs|outputs|both}

Expand globs when storing inputs and/or outputs in the commit message. Constraints: value must be one of ('inputs', 'outputs', 'both')

### -\-explicit

Consider the specification of inputs and outputs to be explicit. Don't warn if the repository is dirty, and only save modifications to the listed outputs.

### -\-sidecar {yes|no}

By default, the configuration variable 'datalad.run.record-sidecar' determines whether a record with information on a command's execution is placed into a separate record file instead of the commit message (default: off). This option can be used to override the configured behavior on a case-by-case basis. Sidecar files are placed into the dataset's '.datalad/runinfo' directory (customizable via the 'datalad.run.record-directory' configuration variable). Constraints: value must be NONE or value must be convertible to type bool

### -\-version

show the module and its version which provides the command

### Authors

datalad is developed by The DataLad Team and Contributors <team@datalad.org>.

## 2.2 Python API

| | |
|---|---|
| *containers_add* | Add a container environment to a dataset |
| *containers_remove* | Remove a container environment from a dataset |
| *containers_list* | List known container environments of a dataset |
| *containers_run* | Drop-in replacement for *datalad run* for command execution in a container |
| *utils* | Collection of common utilities |

### 2.2.1 datalad_container.containers_add

Add a container environment to a dataset

**class** datalad_container.containers_add.**ContainersAdd**

Bases: Interface

Add a container to a dataset

### 2.2.2 datalad_container.containers_remove

Remove a container environment from a dataset

**class** datalad_container.containers_remove.**ContainersRemove**

Bases: Interface

Remove a known container from a dataset

This command is only removing a container from the committed Dataset configuration (configuration scope branch). It will not modify any other configuration scopes.

This command is *not* dropping the container image associated with the removed record, because it may still be needed for other dataset versions. In order to drop the container image, use the 'drop' command prior to removing the container configuration.

### 2.2.3 datalad_container.containers_list

List known container environments of a dataset

**class** datalad_container.containers_list.**ContainersList**

Bases: Interface

List containers known to a dataset

**static custom_result_renderer**(*res*, ***kwargs*)

**result_renderer = 'tailored'**

### 2.2.4 datalad_container.containers_run

Drop-in replacement for *datalad run* for command execution in a container

`class datalad_container.containers_run.`**`ContainersRun`**

Bases: `Interface`

Drop-in replacement of 'run' to perform containerized command execution

Container(s) need to be configured beforehand (see containers-add). If no container is specified and only one container is configured in the current dataset, it will be selected automatically. If more than one container is registered in the current dataset or to access containers from subdatasets, the container has to be specified.

A command is generated based on the input arguments such that the container image itself will be recorded as an input dependency of the command execution in the *run* record in the git history.

During execution the environment variable {name_envvar} is set to the name of the used container.

**`on_failure = 'stop'`**

### 2.2.5 datalad_container.utils

Collection of common utilities

`datalad_container.utils.`**`get_container_command`**`()`

`datalad_container.utils.`**`get_container_configuration`**`(`*ds: Dataset*, *name: str | None = None*`)` → dict

Report all container-related configuration in a dataset

Such configuration is identified by the item name pattern:

```
datalad.containers.<container-name>.<item-name>
```

**Parameters**

- **ds** (*Dataset*) -- Dataset instance to report configuration on.
- **name** (*str, optional*) -- If given, the reported configuration will be limited to the container with this exact name. In this case, only a single `dict` is returned, not nested dictionaries.

**Returns**

Keys are the names of configured containers and values are dictionaries with their respective configuration items (with the `datalad.containers.<container-name>.` prefix removed from their keys). If *name* is given, only a single `dict` with the configuration items of the matching container is returned (i.e., there will be no outer `dict` with container names as keys). If not (matching) container configuration exists, and empty dictionary is returned.

**Return type**

dict

# PYTHON MODULE INDEX

# INDEX