o2r opening reproducible research

Automatically Archiving Reproducible Studies with Docker

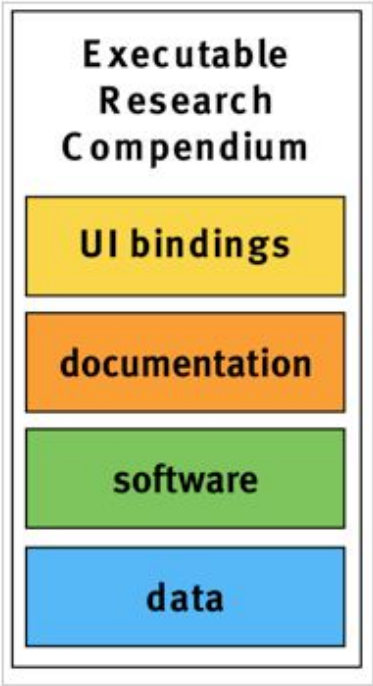
Daniel Nüst | University of Münster | @nordholmen

useR!2017 Brussels, Thu 12:12 pm, room 2.02

http://sched.co/AxqM

Contents

#1: reproducibility helps to avoid disaster
#2: reproducibility makes it easier to write papers
#3: reproducibility helps reviewers see it your way
#4: reproducibility enables continuity of your work
#5: reproducibility helps to build your reputation

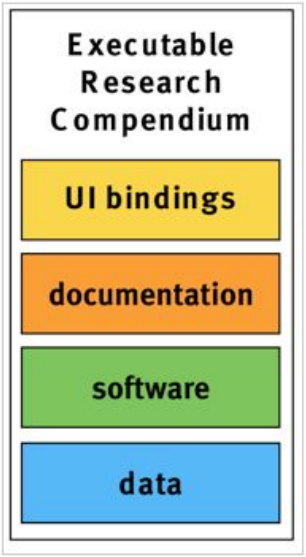


Opening the Publication Process with Executable Research Compendia *Nüst D, Konkol M, Pebesma E, Kray C, Schutzeichel M, Przibytzin, H, Lorenz, J*
Article(journal). D-Lib Magazine. 2017. **doi: 10.1045/january2017-nuest**;
http://www.dlib.org/dlib/january17/nuest/01nuest.html

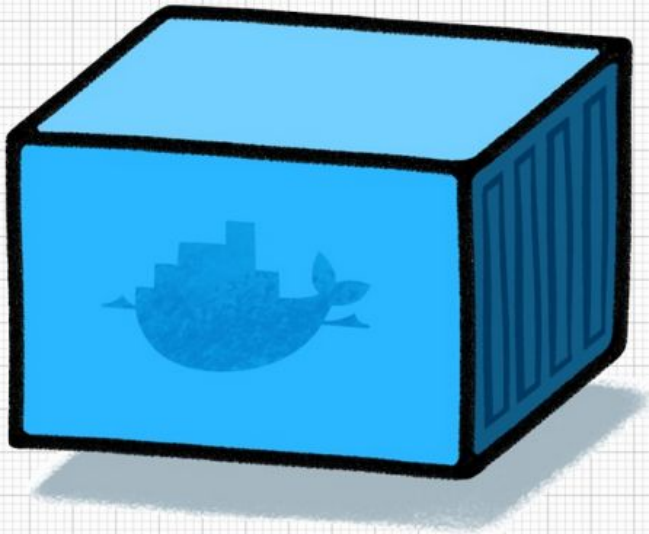
o2r opening reproducible research

ERC creation process

- Submit workspace to **publication platform**
- Publication platform…
  - extracts metadata
  - executes analysis
    - check output vs. upload (syntax)
    - **capture runtime environment (manifest + image)**
  - User checks metadata
  - Publish ERC persistently



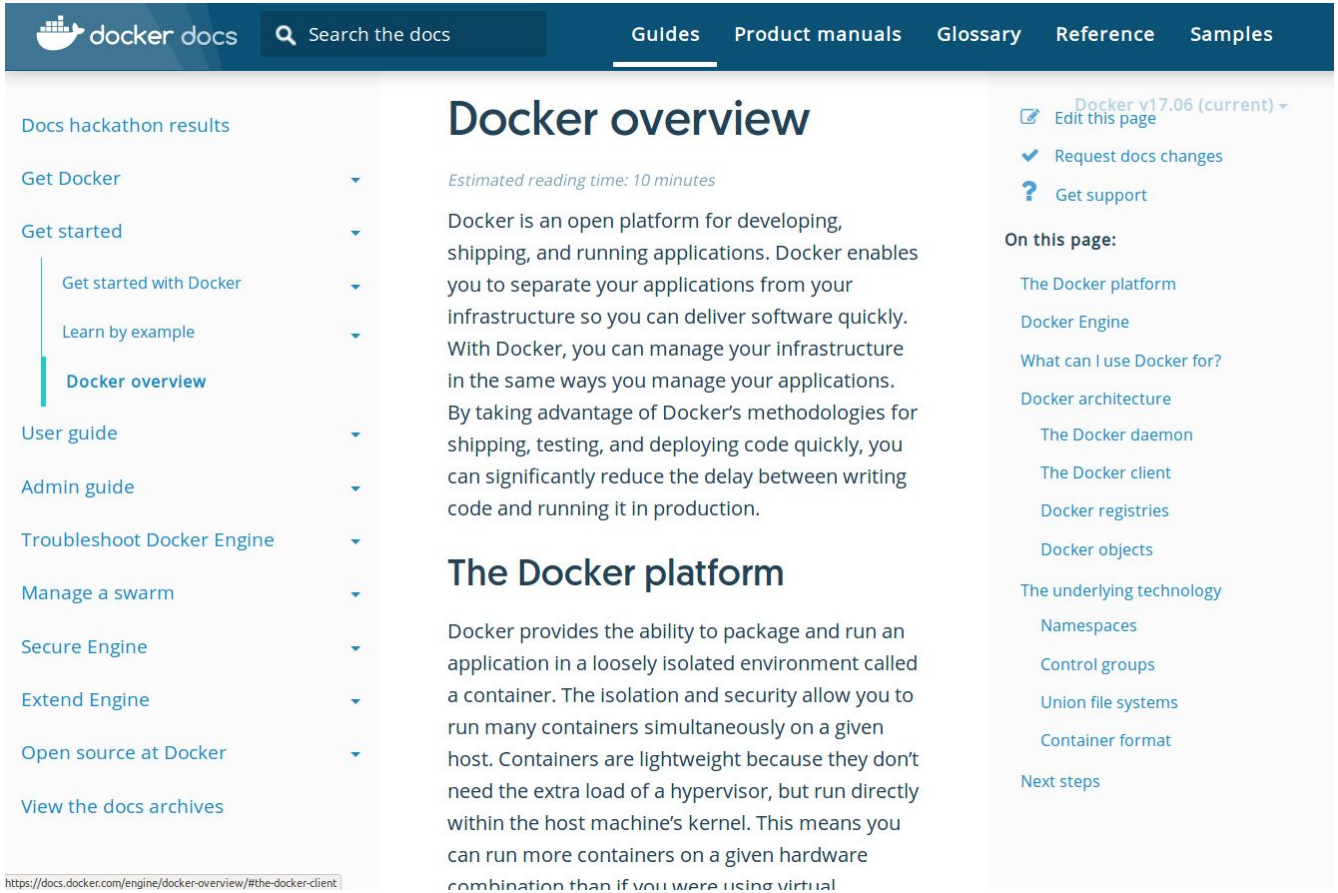
The real value of Docker is not technology



It's getting people to agree on something

Slide by Docker inventor & Docker, Inc. CTO Solomon Hykes, DockerCon 2014

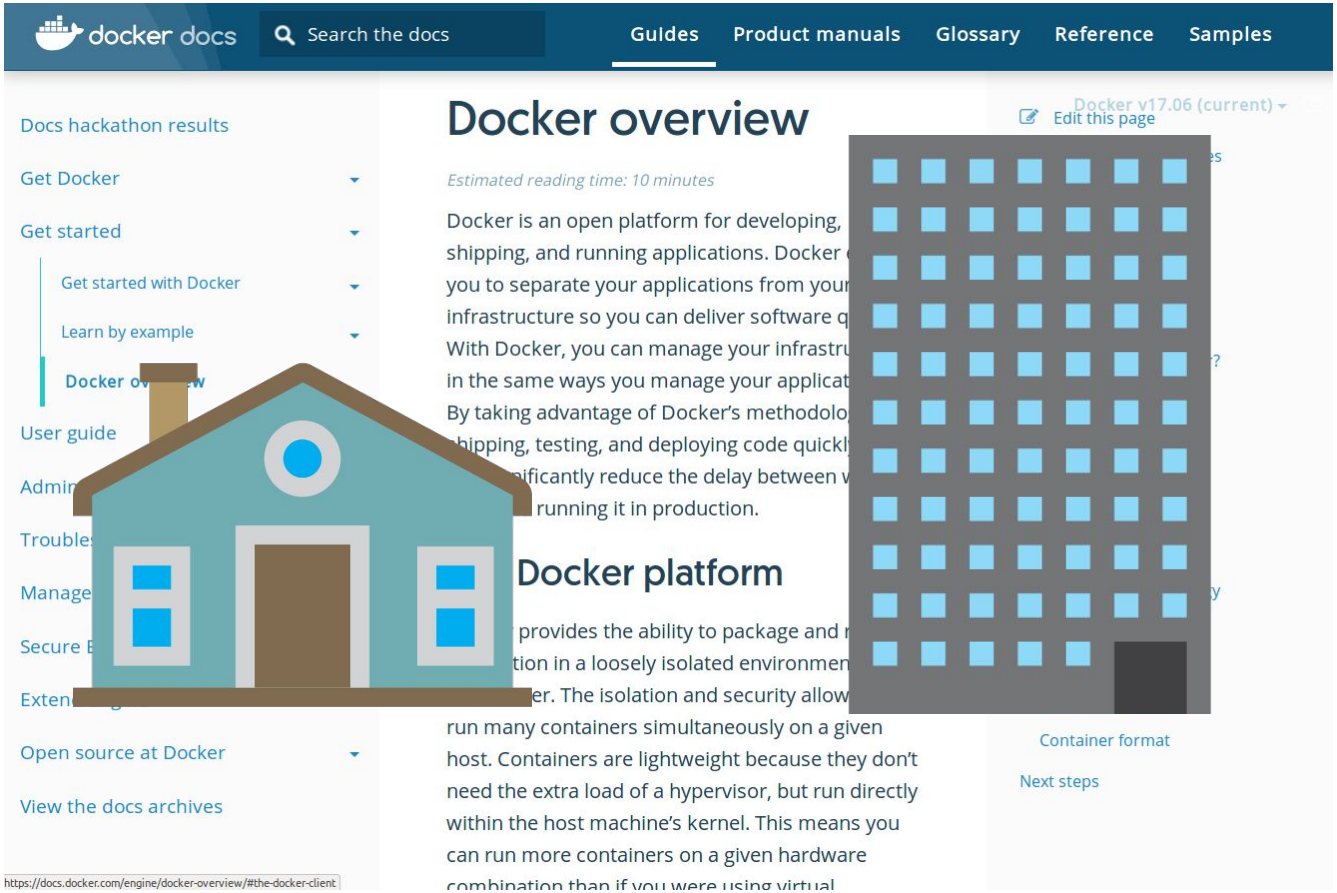
giphy.com



~~science~~
~~data science~~
~~research~~
~~reproducibility~~
~~replication~~

package & separate applications and their dependencies for cloud infrastructures

https://docs.docker.com/engine/docker-overview/

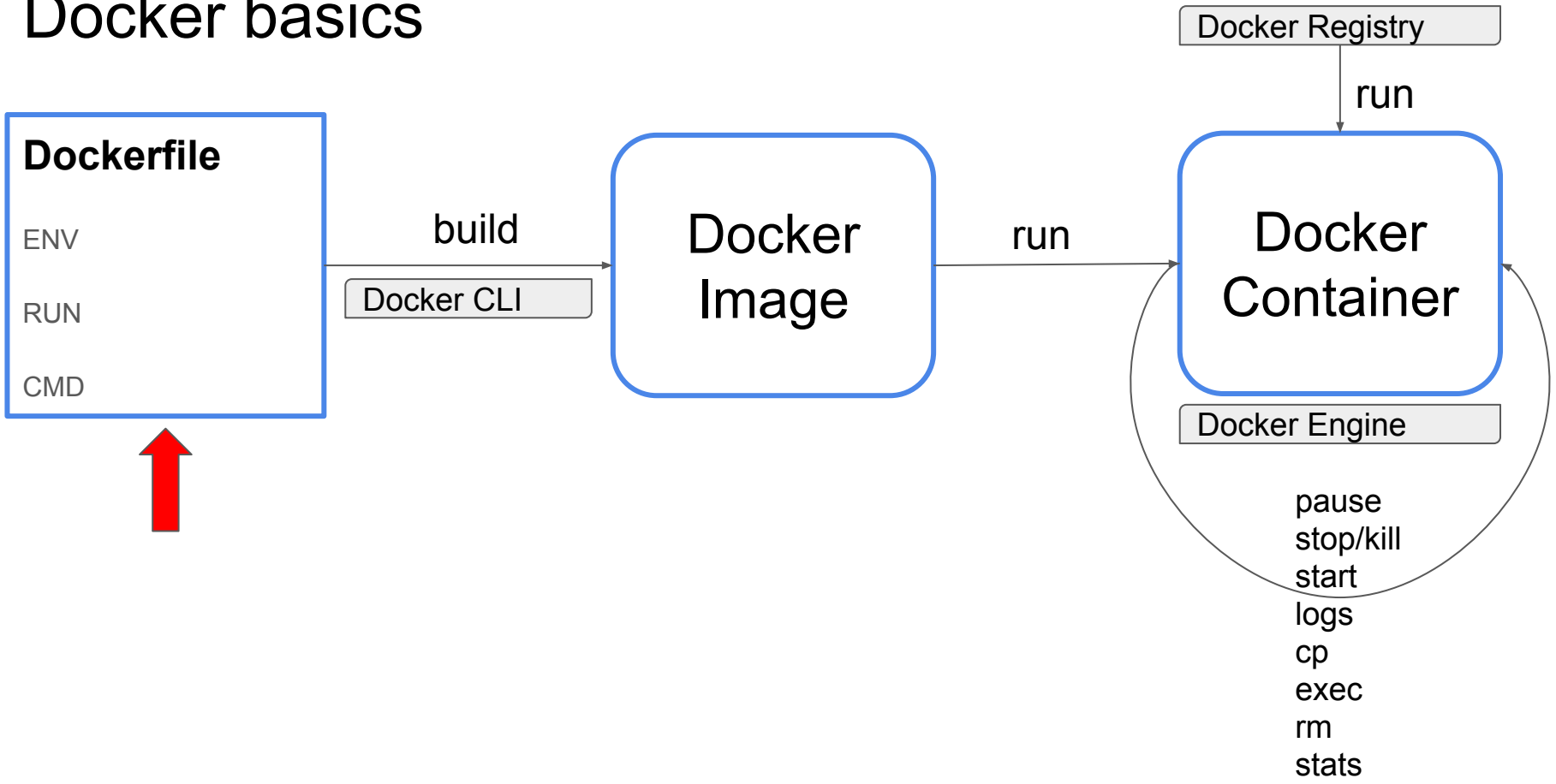


~~science~~
~~data science~~
~~research~~
~~reproducibility~~
~~replication~~

package & separate applications and their dependencies for cloud infrastructures

https://docs.docker.com/engine/docker-overview/

Docker basics



Docker for Data Science

(all the Docker advantages… write once, biz ops, cloud, etc.)

Reproducibility

Project separation + don’t clutter dev machine

Environment (re)creation, **documentation**

Adopt **good practices** on the way

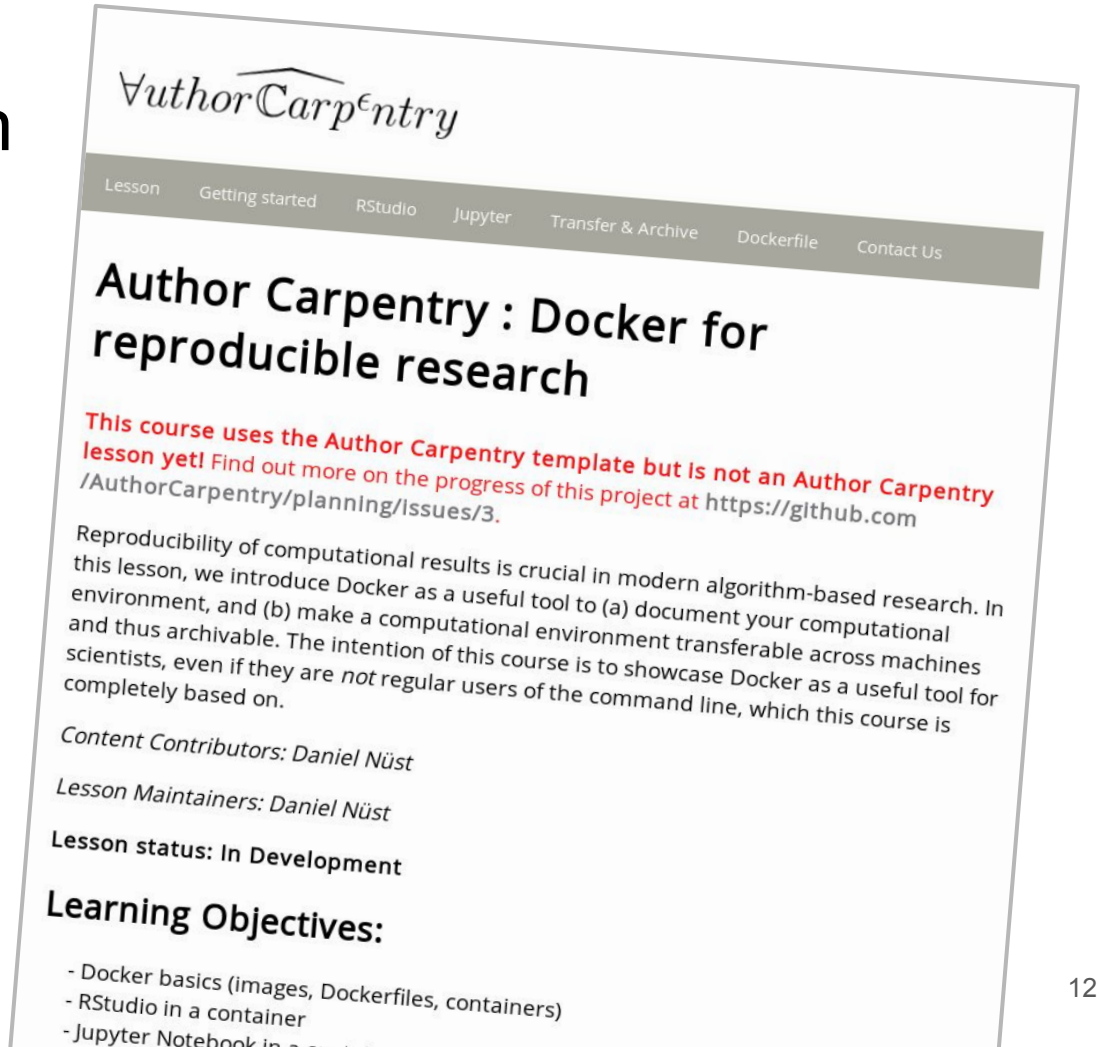
Easy **collaboration**

Easy **transition** from testing to production

Docker for RR lesson

https://github.com/nuest/docker-reproducible-research

https://nuest.github.io/docker-reproducible-research



Rocker: https://github.com/rocker-org

https://hub.docker.com/r/rocker/rstudio/

Base containers (r-base, r-devel, **r-ver**, ..)

Use case containers (r-devel-ubsan-clang, ..)

Stacks (tidyverse, geospatial, ..)

```
docker run -it -p 8787:8787 rocker/rstudio
```

http://localhost:8787/ (rstudio/rstudio)

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

https://bioconductor.org/help/docker/

Current Containers

Maintained by the Bioconductor Core Team: bioc-issue-bot@bioconductor.org

- bioconductor/devel_base2
- bioconductor/devel_core2
- bioconductor/release_base2
- bioconductor/release_core2

Maintained by Steffen Neumann: sneumann@ipb-halle.de

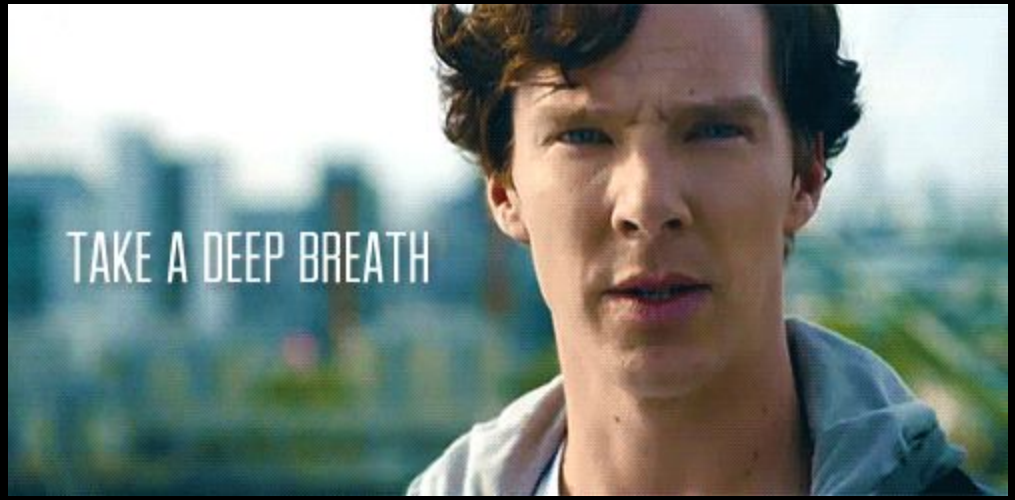
Maintained as part of the "PhenoMeNal, funded by Horizon2020 grant 654241"

- bioconductor/devel_protmetcore2
- bioconductor/devel_metabolomics2
- bioconductor/release_protmetcore2
- bioconductor/release_metabolomics2

Maintained by Laurent Gatto: lg390@cam.ac.uk

- bioconductor/devel_proteomics2
- bioconductor/release_proteomics2

https://hub.docker.com/u/bioconductor/

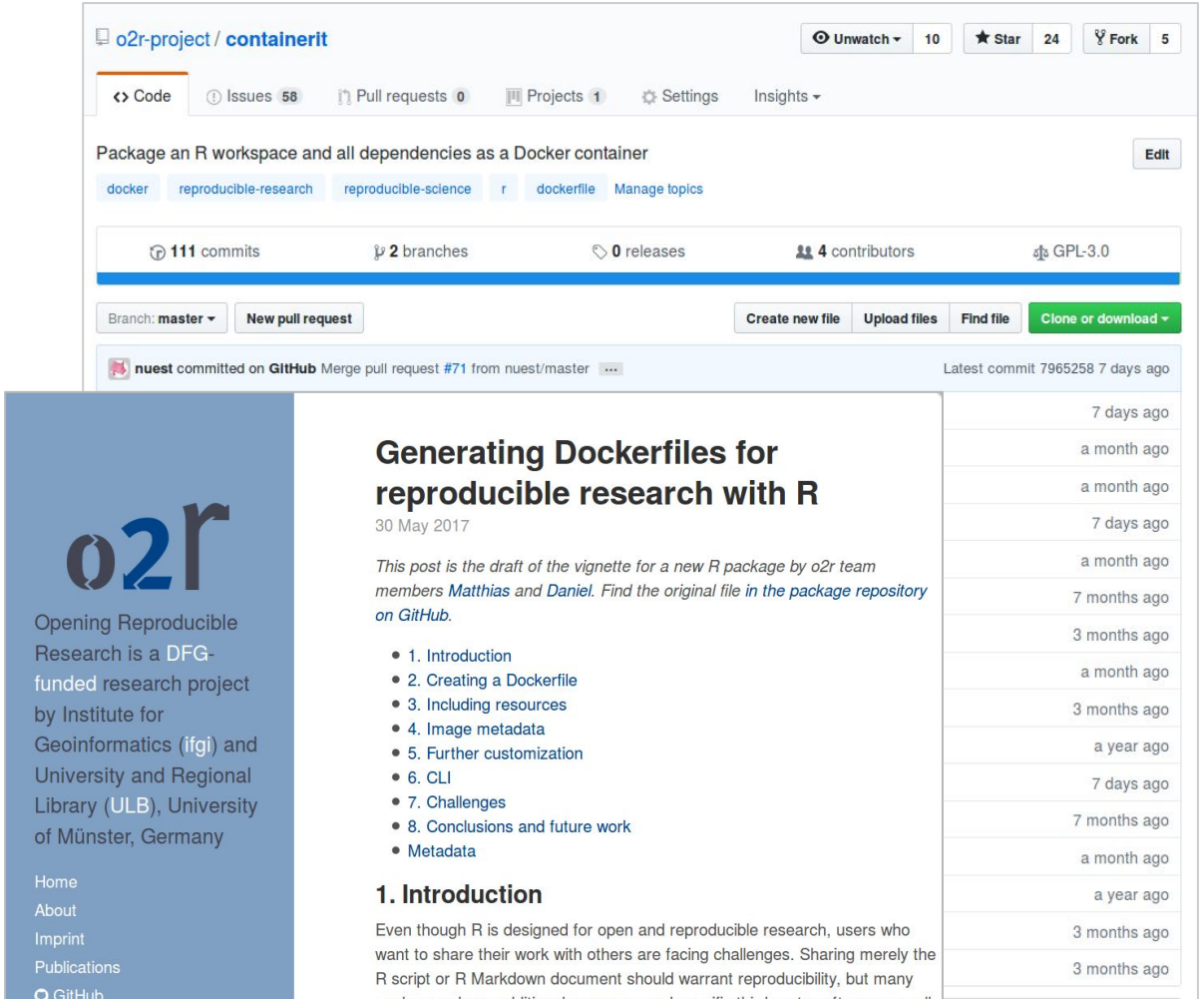


giphy.com

containerit

https://github.com/o2r-project/containerit

http://o2r.info/2017/05/30/containerit-package/



Packaging interactive session

```
> library(containerit); library("gstat"); library("sp")
> data(meuse)
> coordinates(meuse) = ~x+y
> data(meuse.grid)
> gridded(meuse.grid) = ~x+y
> v <- variogram(log(zinc)~1, meuse)
> m <- fit.variogram(v, vgm(1, "Sph", 300, 1))
> plot(v, model = m)
```

```
> dockerfile_object <- dockerfile()
```

```
INFO [2017-07-05 11:20:54] Trying to determine system
requirements for the package(s) 'sp, gstat, zoo,
futile.logger, xts, lambda.r, spacetime, futile.options,
FNN, intervals, lattice' from sysreq online DB
INFO [2017-07-05 11:21:03] Adding CRAN packages: sp,
gstat, zoo, futile.logger, xts, lambda.r, spacetime,
futile.options, FNN, intervals, lattice
INFO [2017-07-05 11:21:03] Created Dockerfile-Object
based on sessionInfo
```

```
> print(dockerfile_object)
```

```
FROM rocker/r-ver:3.4.1
LABEL maintainer="daniel"
RUN ["install2.r", "-r 'https://cloud.r-project.org'", "sp",
"gstat", "zoo", "futile.logger", "xts", "lambda.r",
"spacetime", "futile.options", "FNN", "intervals",
"lattice"]
WORKDIR /payload/
CMD ["R"]
```

```
> str(dockerfile_object, max.level = 2)
Formal class 'Dockerfile' [..] with 4 slots
  ..@ image        :Formal class 'From' [..] with 2 slots
  ..@ maintainer   :Formal class 'Label' [..] with 2 slots
  ..@ instructions :List of 2
  ..@ cmd          :Formal class 'Cmd' [..] with 2 slots
```

Packaging a script w/ sysreqs dependency resolving

```
library(rgdal); require(maptools)
nc <- rgdal::readOGR(system.file("shapes/",
package="maptools"), "sids", verbose = FALSE)
proj4string(nc) <- CRS("+proj=longlat
+datum=NAD27")
plot(nc)
summary(nc)

> scriptCmd <- CMD_Rscript("demo.R")

> dockerfile_object <- dockerfile(
      from = "~/Documents/2017_useR/demo.R",
      cmd = scriptCmd)

# curl https://sysreqs.r-hub.io/pkg/
rgdal,sp,lattice/linux-x86_64-debian-gcc
# ["libgdal-dev", "libproj-dev", "gdal-bin"]
```

```
> print(dockerfile_object)

FROM rocker/r-ver:3.4.0
LABEL maintainer="daniel"
RUN export DEBIAN_FRONTEND=noninteractive; apt-get
-y update \
 && apt-get install -y gdal-bin \
    libgdal-dev \
    libproj-dev
RUN ["install2.r", "-r
'https://cloud.r-project.org'", "rgdal", "sp",
"lattice"]
WORKDIR /payload/
COPY [".", "."]
CMD ["R", "--vanilla", "-f",
"containerit_1a977e2dcdea.R"]
```

Running the container

```
> write(dockerfile_object)
INFO [2017-07-06 10:10:05] Writing dockerfile to
/home/daniel/Documents/2017_useR/Dockerfile
```

```
$ docker build -t user2017demo .
```

```
Sending build context to Docker daemon  6.054MB
Step 1/7 : FROM rocker/r-ver:3.4.1
3.4.1: Pulling from rocker/r-ver
c75480ad9aaf: Pull complete
[...]
The following additional packages will be installed:
[...]
* installing *source* package ‘foreign’ ...
[...]
Successfully built e30936ac8687
Successfully tagged user2017demo:latest
```

```
$ docker run -it user2017demo
```

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical
Computing
Platform: x86_64-pc-linux-gnu (64-bit)
[..]
> library(rgdal); require(maptools)
Loading required package: sp
> nc <- rgdal::readOGR(system.file("shapes/",
package="maptools"), "sids", verbose = FALSE)
[...]
> summary(nc)
Object of class SpatialPolygonsDataFrame
Coordinates:
      min         max
x -84.32385 -75.45698
y  33.88199  36.58965
Is projected: FALSE
[...]
```

CLI

Based on docopt

https://github.com/docopt/docopt.R

```
daniel@gin-nuest:~/git/o2r$ containerit --help
container_it.R is a command-line interface to the R package containerit.
It packages R sessions, scripts, workspaces and vignettes together with all dependencies to ex

Usage:  container_it.R dir [options] [--copy arg] [-d <DIR>]
        container_it.R file [options] [--copy arg] [--cmd-render <FORMAT> | --cmd-R-file] <FIL
        container_it.R session [options] [-e <EXPR> ...]
        container_it.R [--help | --version]

Modes:
    dir       Searches the given directory for R / Rmarkdown files and uses the first encounte
    file      Packages a given R script or R markdown file.
              Optionally, the container can run a batch execution or rendering of the given fi
    session   Packages an empty R session in which a series of R commands may be executed (see

Options (for all modes):
  --force -f           Force writing output even if a file of the same name already exists
  --image <ARG>        Specify the Docker image that shall be used for the docker container (FR
                       By default, the image is determinded from the given r_version,
                       while the version is matched with tags from the base image rocker/r-ver
                       see details about the rocker/r-ver at https://hub.docker.com/r/rocker/r-
  --maintainer -m <ARG>   Name / email of the dockerfile's maintainer (will be read from envir
  --no-write           Don't write dockerfile to output file
  --no-vanilla         Package a session / file without using the vanilla flag
                       (warning: site and environment files currently cannot be included in the
  --output -o FILE     Path and name of the output Dockerfile [default: ./Dockerfile]
  --print -p           Print dockerfile to the console
  --r_version -r <ARG>  Specify an R version number that should run inside the container.
                       By default, the version of the currently linked R instance is used.
  --save -s <objects> ...    Save a list of objects from the workspace to an .RData file (over
  --save-image -i      Save the current workspace to an .RData file
  --soft               Whether to include soft dependencies among the system dependencies of R
  --quiet -q           Run containerit as silent as possible (print only errors and warnings) [

Other:
  --copy -c script | script_dir | <copy_file> ...
                       Indicates whether and how a workspace should be copied.
                       For the modes 'dir' and 'file' containerit copies either the given input
```

CLI

```
daniel@gin-nuest:~/git/o2r/containeRit/tests/scripts$ containerit file -f sf.R
INFO [2017-05-15 17:38:36] Executing R script file in sf.R locally.
INFO [2017-05-15 17:38:36] Creating an R session with the following arguments:
         R  --silent --vanilla -e "source(file = \"/home/daniel/git/o2r/containe
sessionInfo()" -e "save(list = \"info\", file = \"/tmp/RtmpkVl9Kw/rdata-sessioni
> source(file = "/home/daniel/git/o2r/containeRit/tests/scripts/sf.R", echo = TR

> cat("Hello from containerit!\n")
Hello from containerit!

> library("sf")
Linking to GEOS 3.5.1, GDAL 2.1.3, proj.4 4.9.2, lwgeom 2.3.2 r15302

> demo("meuse_sf", ask = FALSE)


        demo(meuse_sf)
        ---- ~~~~~~~~

> data(meuse, package = "sp") # load data.frame from sp
```

▪ ▪ ▪

CLI

```
INFO [2017-05-15 17:38:37] Docker will try to install GDAL 2.1.3 from source
INFO [2017-05-15 17:38:37] Trying to determine system requirements for the package(s) 'sf, magrittr, DBI, units
m sysreq online DB
INFO [2017-05-15 17:38:38] Created Dockerfile-Object based on sf.R
INFO [2017-05-15 17:38:38] Writing dockerfile to ./Dockerfile
daniel@gin-nuest:~/git/o2r/containeRit/tests/scripts$ cat Dockerfile
FROM rocker/r-ver:3.4.0
LABEL maintainer="daniel"
RUN export DEBIAN_FRONTEND=noninteractive; apt-get -y update \
 && apt-get install -y gdal-bin \
        libgeos-dev \
        libproj-dev \
        libudunits2-dev \
        make \
        wget
WORKDIR /tmp/gdal
RUN wget http://download.osgeo.org/gdal/2.1.3/gdal-2.1.3.tar.gz \
 && tar zxf gdal-2.1.3.tar.gz \
 && cd gdal-2.1.3 \
 && ./configure \
 && make \
 && make install \
 && ldconfig \
 && rm -r /tmp/gdal
RUN ["install2.r", "-r 'https://cloud.r-project.org'", "sf", "magrittr", "DBI", "units", "Rcpp", "udunits2"]
WORKDIR /payload/
COPY ["sf.R", "sf.R"]
CMD ["R"]
```

More

Labels for metadata

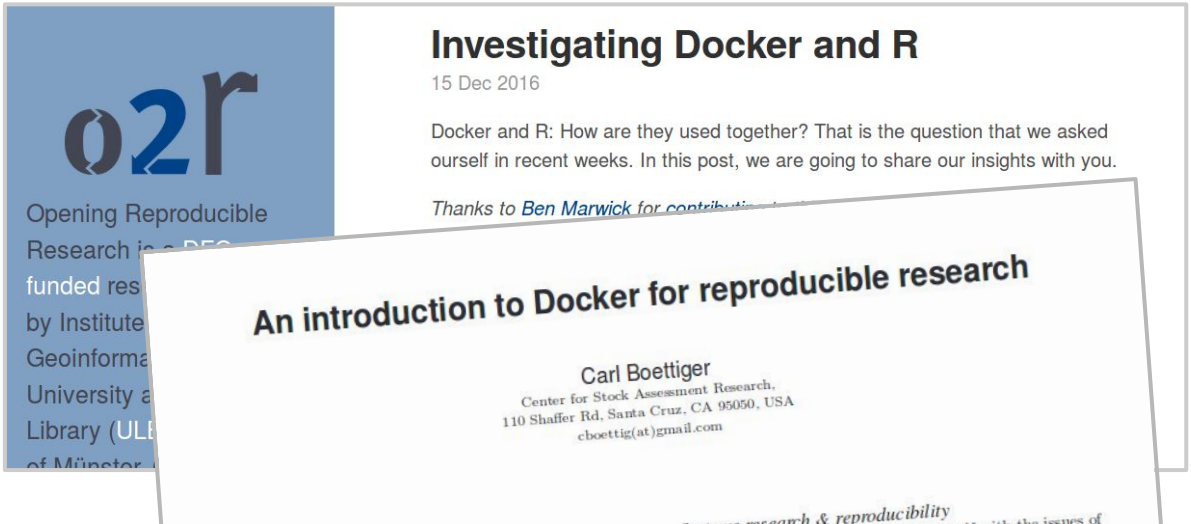
`devtools` session information (install from git under dev.)

Custom **base images**

Docker vs. R blog

http://bit.ly/docker-r

Boettiger, Carl. 2015. “An Introduction to Docker for Reproducible Research, with Examples from the R Environment.” *ACM SIGOPS Operating Systems Review* 49 (January): 71–79. **doi:**10.1145/2723872.2723882



Summary

Executable Research Compendia are fun

Docker is great

`containerit` makes Docker easier BETA!
(DRY, less copy&paste, best practices, automatic **system dependencies**)

Benefits from **Rocker** (MRAN by default, …), harbor, ...

Alternatives / potential for combination:

package management locally (**packrat**, **pkgsnap**, **switchr**/**GRANBase**) or remotely (*MRAN timemachine*/**checkpoint**), or install specific versions from CRAN or source (**requireGitHub**, **devtools**)

Outlook

Support our **ERC creation service**

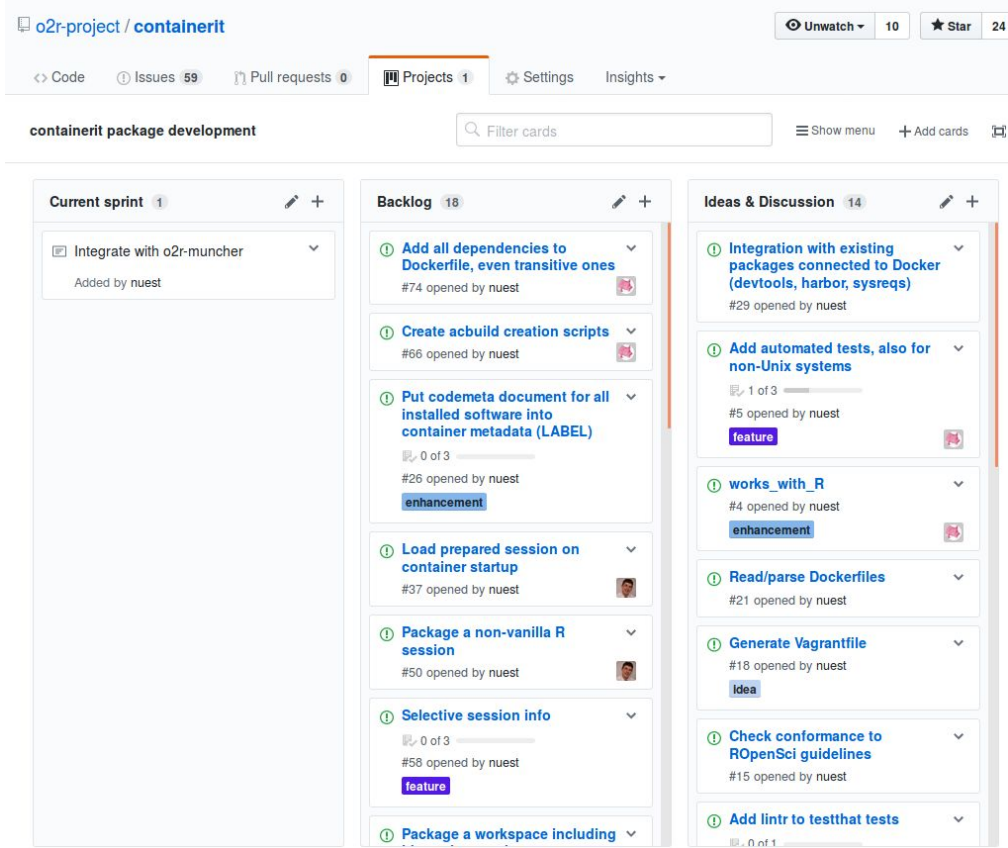
Get feedback

Singularity
OCI/acbuild

CRAN

Docker + R paper for RJournal?

Package rplumber/jug **web apps**



https://github.com/o2r-project/containerit/projects/1

Versioned packages and system libs (`sf::sf_extSoftVersion()`)

giphy.com

Thanks!
What are your questions?

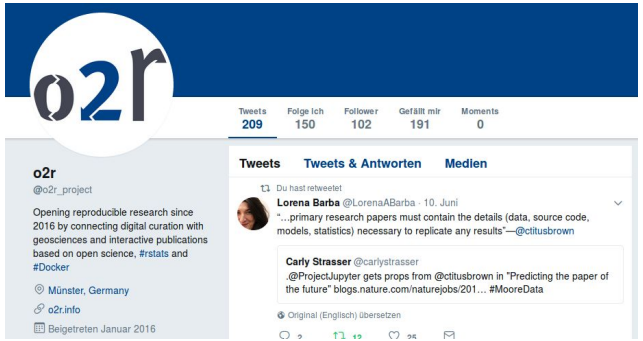
This is the o2r team and supporting university staff in alphabetical order.

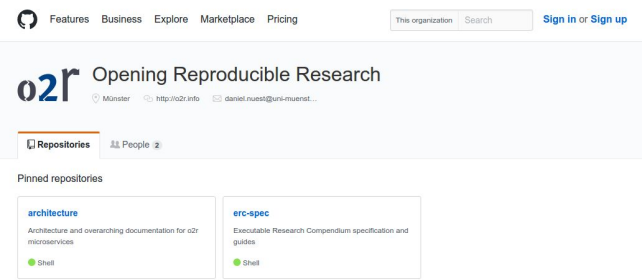
- Rehan Chaudhary (ifgi, internship from 2017-01-17 to 2017-07-17)
- Matthias Hinz (ifgi, 2016-12 to 2017-03)
- Jim Jones (ULB)
- Dr. Stephanie Klötgen (ULB)
- Markus Konkol (ifgi)
- Jan Koppe (ifgi, student assistant, 2016-03 to 2016-08)
- Torben Kraft (ifgi, student assistant)
- Prof. Dr. Christian Kray (ifgi)
- Dr. Dirk Kussmann (ULB)
- Timm Kühnel (ifgi, student assistant)
- Lukas Lohoff (ifgi, student assistant)
- Jörg Lorenz (ULB)
- Daniel Nüst (ifgi)
- Prof. Dr. Edzer Pebesma (ifgi)
- Holger Przibytzin (ULB)
- Dr. Marc Schutzeichel (ULB)
- Jan Suleiman (ifgi, student assistant)
- Dr. Beate Tröger (ULB)

ifgi
Institute for Geoinformatics
University of Münster

ulb Münster

DFG Deutsche Forschungsgemeinschaft





@o2r_project

github.com/o2r-project

o2r.info