# Computational reproducibility with GNU Guix: application to multiparametric cytometry

Simon Tournier<sup>1</sup> & Nicolas Vallet<sup>2</sup>

simon.tournier@u-paris.fr nicolas.vallet@inserm.fr

iPOP-UP, September 9th, 2021

<sup>1</sup>Research Eng. Sci. Computing (+little admin-sys)
<sup>2</sup>PhD student (and MD)



Science = Transparency

Scientific result = Experiment + Numerical treatment

#### Science at the numerical age:

1. Open Article HAL, BioArxiv

2. Open Data Zenodo

**3.** Open Source GitHub, Software Heritage

and how to glue all that?

```
reproductibility = verification replicability = validation
```

R. Di Cosmo@Scibian2016 (pdf) K. Hinsen@Aramis2019 (pdf)

```
Science = Transparency
Scientific result = Experiment + <u>Numerical treatment</u>
```

#### Science at the numerical age:

- Open Article HAL, BioArxiv
- 2. Open Data Zenodo
- 3. Open Source GitHub, Software Heritage
  - . Computational env.

and how to glue all that?

```
reproductibility = verification )
replicability = validation
```

R. Di Cosmo@Scibian2016 (pdf)
K. Hinsen@Aramis2019 (pdf)

|        |              | audit              |   | opaque |   | depend?                               |
|--------|--------------|--------------------|---|--------|---|---------------------------------------|
| result | $\leftarrow$ | paper              | + | data   | + | analysis                              |
|        |              | protocol<br>script |   |        |   | materials (reagent, etc.) environment |

- audit is the "easy" part
- opaque is generally the hard part

|             |        |              | audit              |   | opaque |   | depend?                               |
|-------------|--------|--------------|--------------------|---|--------|---|---------------------------------------|
|             | result | $\leftarrow$ | paper              | + | data   | + | analysis                              |
| $\triangle$ |        |              | protocol<br>script |   |        |   | materials (reagent, etc.) environment |

- audit is the "easy" part
- opaque is generally the hard part

|             |        |              | audit |   | opaque              |   | depend?                               |
|-------------|--------|--------------|-------|---|---------------------|---|---------------------------------------|
|             | result | $\leftarrow$ | paper | + | data                | + | analysis                              |
| $\triangle$ |        |              | •     |   | instruments<br>data |   | materials (reagent, etc.) environment |

- audit is the "easy" part
- opaque is generally the hard part
- how to evacuate depend? from the equations

|             |        |              | audit              |   | opaque |    | depend?                               |
|-------------|--------|--------------|--------------------|---|--------|----|---------------------------------------|
|             | result | $\leftarrow$ | paper              | + | data   | +  | analysis                              |
| $\triangle$ |        |              | protocol<br>script |   |        | ++ | materials (reagent, etc.) environment |

- audit is the "easy" part
- opaque is generally the hard part
- how to evacuate depend? from the equations...

... at least let speak about environment

#### Concrete scenarii about environment

- ▶ Alice used numerical tools R@4.1.1, FlowCore@2.4 and CATALYST@1.16.2.
- ► Carole <u>collaborates</u> with Alice ...

  but also requires R@4.0.4. FlowCore@2.0 for another project.
- Charlie upgrades their system then all is broken.
- ▶ Bob runs the same versions as Alice ...

but does not get the same outputs.

▶ Dan tries to redo the analysis months (years?) later . . .

but hits the dependencies hell<sup>1</sup>.

<sup>1</sup>at best :-)

#### Concrete scenarii about environment

- ▶ Alice used numerical tools R@4.1.1, FlowCore@2.4 and CATALYST@1.16.2.
- ► Carole collaborates with Alice ...

but also requires R@4.0.4, FlowCore@2.0 for another project.

- Charlie upgrades their system then all is broken.
- ▶ Bob runs the same versions as Alice ...

but does not get the same outputs.

▶ Dan tries to <u>redo</u> the analysis months (years?) later . . .

but hits the dependencies hell<sup>1</sup>.

#### Solution(s)

- 1. package manager: apt, yum, etc.
- 2. virtual environment: conda, modulefiles, etc.
- **3. container**: Docker, Singularity, etc.

Guix

1at best :-)

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

What is the issue? **R is Open Source** https://svn.r-project.org/R/trunk/src/main/main.c

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

Recipe to make an yogurt:

```
Yogurt ← Milk + Skyr<sup>2</sup>
```

<sup>&</sup>lt;sup>2</sup>Icelandic strained vogurt

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

What is the issue? **R is Open Source** https://svn.r-project.org/R/trunk/src/main/main.c

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

Recipe to make an yogurt executable:

<sup>&</sup>lt;sup>2</sup>Icelandic strained vogurt

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

What is the issue? **R is Open Source** https://svn.r-project.org/R/trunk/src/main/main.c

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

Recipe to make an yogurt executable:

<sup>&</sup>lt;sup>2</sup>Icelandic strained vogurt

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

What is the issue? **R is Open Source** https://svn.r-project.org/R/trunk/src/main/main.c

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

Recipe to make an yogurt executable:

```
Yogurt ← Milk + Skyr²

executable source

R main.c C compiler

executable ← source + executable
```

<sup>&</sup>lt;sup>2</sup>Icelandic strained yogurt

```
alice@laptop$ R -e '1+2'
> 1+2
[1] 3
```

```
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
```

```
alice@laptop$ R -e '1+2'
    > 1+2
    Γ1 3
                   What is the issue? R is Open Source
              https://svn.r-project.org/R/trunk/src/main/main.c
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
    alice@laptop$ ldd /usr/lib/R/bin/exec/R
        libblas.so.3 => /usr/lib/x86_64-linux-gnu/libblas.so.3
        libR.so => /usr/lib/libR.so
        libgfortran.so.5 => /usr/lib/x86_64-linux-gnu/libgfortran.so.5
        libquadmath.so.0 => /usr/lib/x86_64-linux-gnu/libquadmath.so.0
```

```
alice@laptop$ R -e '1+2'
    > 1+2
    Γ1 3
                   What is the issue? R is Open Source
              https://svn.r-project.org/R/trunk/src/main/main.c
alice@laptop$ file /usr/lib/R/bin/exec/R
/usr/lib/R/bin/exec/R: executable, dynamically linked, ...
    alice@laptop$ ldd /usr/lib/R/bin/exec/R
        libblas.so.3 => /usr/lib/x86_64-linux-gnu/libblas.so.3
        libR.so => /usr/lib/libR.so
        libgfortran.so.5 => /usr/lib/x86_64-linux-gnu/libgfortran.so.5
        libquadmath.so.0 => /usr/lib/x86_64-linux-gnu/libquadmath.so.0
```

K. Hinsen@Aramis2019 (video)

Blog: Reproducible computations with Guix (link)

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes, but...

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes. but...

Yogurt ← Milk + Skyr + Utensils
executable source executable co)
R main.c C compiler ← Deps. (libblas, etc.)

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes. but...

```
Yogurt ← Milk + Skyr + Utensils
executable
R main.c + Skyr + Utensils
links(executable too)
C compiler Deps. (libblas, etc.)
```

 $R@4.1.1 \leftarrow source@4.1.1 + executable@A + links@C$ 

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes. but...

```
Yogurt ← Milk + Skyr + Utensils
executable source executable main.c + Skyr + Utensils
links(executable too)
C compiler Deps. (libblas, etc.)
```

 $R@4.1.1 \leftarrow source@4.1.1 + executable@A + links@C$ 

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes. but...

```
      Yogurt ← executable R
      Milk source main.c
      + Skyr + Utensils links(executable too)

      C compiler
      Deps. (libblas, etc.)
```

```
R@4.1.1 \leftarrow source@4.1.1 + executable@A + links@C ||?
R@4.1.1 \leftarrow source@4.1.1 + executable@B + links@D
```

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- Yes, but...

```
      Yogurt ← R
      Milk source main.c
      + Skyr + Utensils links(executable too)

      C compiler
      Deps. (libblas, etc.)
```

► Too many combinations to have the same environment on different machines when something is going wrong, it is hard to say from where it comes is it the environment? is it the analysis? something else?

- ▶ Wait! Package managers (apt, yum, conda, brew, etc.) do the job for us.
- ► Yes, but... e.g., Conda is not enough

```
Yogurt ← Milk + Skyr + Utensils
executable source executable main.c + Skyr + Utensils
| C compiler | Deps. (libblas, etc.)
```

```
||?
R@4.1.1 ← source@4.1.1 + executable@B + links@D

Too many combinations to have the same environment on different machines
```

 $R@4.1.1 \leftarrow source@4.1.1 + executable@A + links@C$ 

when something is going wrong, it is hard to say from where it comes

is it the environment? is it the analysis? something else?

Conclusion: Tools allowing to capture the environment (= @4.1.1, @A, @C)

► Alice says she uses R@4.1.1

alice@laptop\$ install r@4.1.1

alice@laptop\$ R --version

R version 4.1.1 (2021-08-10) -- "Kick Things"

- Alice says she uses R@4.1.1 alice@laptop\$ install r@4.1.1 alice@laptop\$ R --version R version 4.1.1 (2021-08-10) -- "Kick Things"
- Bob runs this same version @4.1.1

R version 4.1.1 (2021-08-10) -- "Kick Things"

- bob@cluster\$ install r@4.1.1 bob@cluster\$ R --version

- ► Alice says she uses R@4.1.1

  alice@laptop\$ install r@4.1.1

  alice@laptop\$ R --version

  R version 4.1.1 (2021-08-10) -- "Kick Things"
- R version 4.1.1 (2021-08-10) -- "Kick Things"

  Bob runs this same version @4.1.1
- bob@cluster\$ install r@4.1.1
  bob@cluster\$ R --version
- R version 4.1.1 (2021-08-10) -- "Kick Things"
- but Bob does not get the same result (e.g., precision, performance, other)

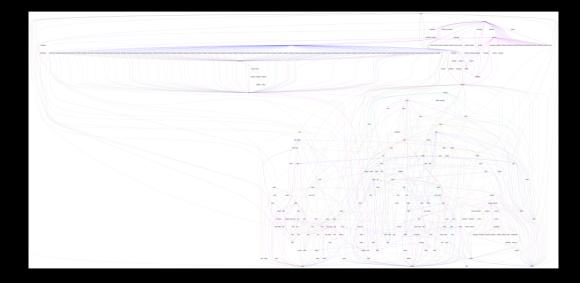
- Alice says she uses R@4.1.1 alice@laptop\$ install r@4.1.1 alice@laptop\$ R --version R version 4.1.1 (2021-08-10) -- "Kick Things"
- Bob runs this same version @4.1.1 bob@cluster\$ install r@4.1.1 bob@cluster\$ R --version
- R version 4.1.1 (2021-08-10) -- "Kick Things"
- but Bob does not get the same result (e.g., precision, performance, other)
- ▶ why? what is different? both uses R@4.1.1

- Alice says she uses R@4.1.1 alice@laptop\$ install r@4.1.1 # gcc-toolchain@7.5.0 alice@laptop\$ R --version R version 4.1.1 (2021-08-10) -- "Kick Things"
- ▶ Bob runs this same version @4.1.1 bob@cluster\$ install r@4.1.1 # gcc-toolchain@9.4.0 bob@cluster\$ R --version R version 4.1.1 (2021-08-10) -- "Kick Things"
- but Bob does not get the same result (e.g., precision, performance, other)
- why? what is different? both uses R@4.1.1
  \$ diff with-{7.5.0.9.4.0}/lib/R/bin/exec/R
  - Binary files differ
  - Some links (libquadmath.so) are at different versions

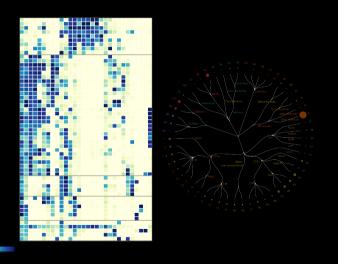
- Alice says she uses R@4.1.1
   alice@laptop\$ install r@4.1.1 # gcc-toolchain@7.5.0
   alice@laptop\$ R --version
   R version 4.1.1 (2021-08-10) -- "Kick Things"
- ▶ Bob runs this same version @4.1.1 bob@cluster\$ install r@4.1.1 # gcc-toolchain@9.4.0 bob@cluster\$ R --version R version 4.1.1 (2021-08-10) -- "Kick Things"
- but Bob does not get the same result (e.g., precision, performance, other)
- ▶ why? what is different? both uses R@4.1.1
  - > \$ diff with-{7.5.0,9.4.0}/lib/R/bin/exec/R
    Binary files differ
- Conclusion: Version @4.1.1 of source is not enough for reproducing

Reproducibility requires to capture the environment

# Concretely, just R requires a lot!



## Usual (soon?) published result



- Cytometry CyTOF:43 parameters
- zillions of events (cells)
- ► FlowJo (cleaning)
- R plus 19 packages from BioConductor

## How to redo this (soon?) published result

#### Assuming nothing is lost and all is transparent

|         |              | audit                                |   | opaque              |   | depend?                  |  |
|---------|--------------|--------------------------------------|---|---------------------|---|--------------------------|--|
| result  | $\leftarrow$ | paper                                | + | data                | + | analysis                 |  |
|         |              |                                      |   | instruments<br>data |   | materials<br>environment |  |
|         |              | environment = collection of packages |   |                     |   |                          |  |
| package | $\leftarrow$ | source                               | + | executable          | + | others packages          |  |

### How to redo this (soon?) published result

#### Assuming nothing is lost and all is transparent

|        |             | audit                                |   | opaque              |   | depend?                  |
|--------|-------------|--------------------------------------|---|---------------------|---|--------------------------|
| result | <del></del> | paper                                | + | data                | + | analysis                 |
|        |             | •                                    |   | instruments<br>data |   | materials<br>environment |
|        |             | environment = collection of packages |   |                     |   |                          |

```
package \leftarrow source + executable (+ others packages)
```

Alice says she uses R@4.1.1 with CATALYST@1.16.2 from Bioconductor means:

### How to redo this (soon?) published result

#### Assuming nothing is lost and all is transparent

|        |              | audit |   | opaque              |   | depend?                  |
|--------|--------------|-------|---|---------------------|---|--------------------------|
| result | $\leftarrow$ | paper | + | data                | + | analysis                 |
|        |              |       |   | instruments<br>data |   | materials<br>environment |

environment = collection of packages

```
package \leftarrow source + executable (+ others packages)
```

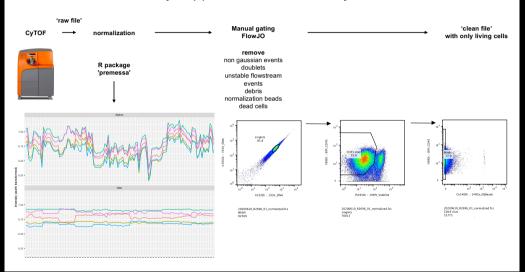
Alice says she uses R@4.1.1 with CATALYST@1.16.2 from Bioconductor means:

- ▶ R: 40 input packages and 40 build packages (but 408 closure packages)
- CATALYST: 27 input packages and 48 build packages

(but 907 closure packages)

# Redo data is complicated

#### CyTOF pipeline before downstream analyses



► Alice lists the packages with the file requirements.scm
(specifications->manifest
(list
 "r"
 "r-catalyst"
 "r-flowcore"))

- ► Alice lists the packages with the file requirements.scm
- Alice describe her current state

```
alice@laptop$ guix describe
Generation 65 Aug 16 2021 20:42:44 (current)
  guix a9eb969
    repository URL: https://git.savannah.gnu.org/git/guix.git
    branch: master
    commit: a9eb969bb63fc421e5cb2a699ef1f1e7c5cbd1b4
alice@laptop$ guix describe -f channels > alice-state.scm
```

- Alice lists the packages with the file requirements.scm
- Alice describe her current state

```
alice@laptop$ guix describe
Generation 65 Aug 16 2021 20:42:44 (current)
  guix a9eb969
    repository URL: https://git.savannah.gnu.org/git/guix.git
    branch: master
    commit: a9eb969bb63fc421e5cb2a699ef1f1e7c5cbd1b4
alice@laptop$ guix describe -f channels > alice-state.scm
```

Bob recreates the exact same environment as Alice

- Alice lists the packages with the file requirements.scm
- Alice describe her current state

```
alice@laptop$ guix describe
Generation 65 Aug 16 2021 20:42:44 (current)
  guix a9eb969
    repository URL: https://git.savannah.gnu.org/git/guix.git
    branch: master
    commit: a9eb969bb63fc421e5cb2a699ef1f1e7c5cbd1b4
alice@laptop$ guix describe -f channels > alice-state.scm
```

▶ Bob recreates the exact same environment as Alice

**Conclusion**: All executable composing the environment are captured by the state.

▶ guix time-machine allows to rebuild the same environment from source

- ▶ guix time-machine allows to rebuild the same environment from source
- Bob runs analysis on another infrastructure

- guix time-machine allows to rebuild the same environment from source
- Bob runs analysis on another infrastructure

This Docker image disappears

- ▶ guix time-machine allows to rebuild the same environment from source
- Bob runs analysis on another infrastructure

#### This Docker image disappears

- ▶ guix time-machine allows to rebuild the same environment from source
- ▶ Bob runs analysis on another infrastructure

#### This Docker image disappears

- Dan runs on another infrastructure using the exact same tools as Alice dan@AWS\$ docker load < image.tar.gz dabn@AWS\$ docker run ...

- ▶ guix time-machine allows to rebuild the same environment from source
- ▶ Bob runs analysis on another infrastructure

```
bob@cluster$ guix time-machine -C alice-state.scm
-- pack -f docker -m requirements.txt
bob@IFB$ docker load < image.tar.gz
bob@IFB$ docker run ...
```

#### This Docker image disappears

- ▶ Dan runs on another infrastructure using the exact same tools as Alice dan@AWS\$ docker load < image.tar.gz dabn@AWS\$ docker run ...
  Lack real examples to ensure this scenario fully works ;-)

A. Legrand@GuixHPC2021 (video)(pdf)

### **Long-term with Software Heritage**

#### What happens if the package on GitHub disappears?

```
dan@workstation$ guix install hi
fatal: could not read Username for 'https://github.com': No such device
or address
Trving content-addressed mirror at berlin.guix.gnu.org...
Trying to download from Software Heritage...
SWH: found revision eleefd033b8a2c4c81babc6fde08ebb116c6abb8
at 'https://archive.softwareheritage.org/api/1/directory/c3e538ed2de412...
The source of hi were at https://github.com/zimoun/hello-example.
ig(See Software Heritage ambassador Pierre Poulain :-ig)
                                                      \left( \, Package defined \underline{\mathsf{here}}.\,\, 
ight)
```

### Guix in a nutshell

### Tools helping in managing complexity

- 1. package manager: functional and transactional paradigm
- 2. environment manager: switch between projects
- 3. generate container: Docker, Singularity
- 4. provide tools (Scheme library) to extend for user specificities

### (fine) Control over the dependency chain

- ★★ Binary reproductibility (at least to track it)
- $\star\star$  Bootstrap (new yogurt  $\leftarrow$  milk + old yogurt)

#### Similar to other package manager

```
# Alice
guix install foo@1.2 bar@3.4 baz@5.6 # -m requirements.scm
guix describe -f channels > alice-conf.scm
# Carole
guix install --profile=./with-alice foo@1.2 bar@3.4 baz@5.6
guix install foo@7.8 bar@9.0
# Charlie
guix package --roll-back
                                           # generations
# Bob
guix pull --channels=alice-conf.scm
guix pack -f docker ...
                                           # -f squashfs
# Dan
guix time-machine -C alice-conf.scm -- install foo@1.2 bar@3.4 baz@5.6
```

## The ideal target



Software Heritage



The Re**Science** Journal



### **Clusters using Guix**

**GriCAD** CCIPL PlaFRIM Inria

Max Delbrück Center

**UMC Utrecht** 

(Univ. Paris?)

Grenoble:

72-node Nantes: 230-node

(1000+ cores)

(4000+ cores) (3000+ cores)

**Bordeaux:** 120-node Berlin: 250-node

68-node

(1.000 + cores)

: ?9-node? + 2 workstations

+ workstations



https://hpc.guix.info

nicolas.vallet@inserm.fr

https://hpc.guix.info



Copyright © 2021 Simon Tournier simon.tournier@u-paris.fr.
Copyright © 2021 Nicolas Vallet nicolas.vallet@inserm.fr.

GNU Guix logo, CC-BY-SA 4.0. https://gnu.org/s/guix/graphics

GNU Guix Reference Card under GFDL 1.3+.
Copyright of other images included in this document is held by their respective owners; especially from Ludovic Courtès.

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 License. To view a copy of this license, visit

http://creativecommons.org/licenses/by-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

At your option, you may instead copy, distribute and/or modify this document under the terms of the GNU Free Documentation License. Version 1.3 or any later version published by the Free Software Foundation: with no Invariant Sections, no

Front-Cover Texts, and no Back-Cover Texts. A copy of the license is available at https://www.gnu.org/licenses/gfdl.html.

The source of this document is available from https://git.sv.gnu.org/cgit/guix/maintenance.git. TODO.

### Container is a smoothie

- ► Container = a binary box that stores (captures) all the executables
- ► How to generate one?

For instance, Dockerfile based on some Linux distribution

What happens if this binary container is lost? Can I rebuild it?

The answer is no, you cannot!

Well, require a lot of hard and low-level work.

A. Legrand@GuixHPC2021 (video)(pdf)

**Conclusion 1**: Container is **useful** to move binaries from one place to another

Conclusion 2: Container is **not the solution** for the Reproducibility Crisis (= run the same analysis on different machines at different moments)