Enhancing Scholarly Communication with ReproZip

Fernando Chirigati, Rémi Rampin, Victoria Steeves, Dennis Shasha, and Juliana Freire

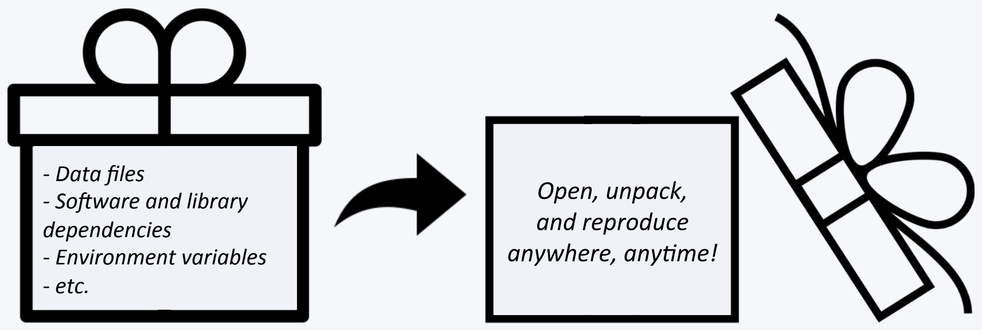
NYU

Reproducibility is Necessary, but Hard ...

... Until now!

ReproZip!

Pack your experiment on your system ***S*** ...
... and **unpack** on another system ***S'*** ...
... using as few as 2 commands for each step !

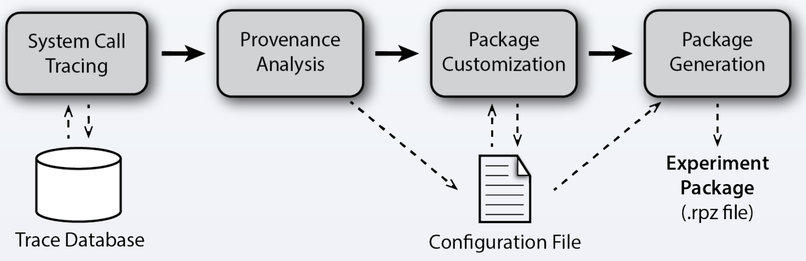


reprozip trace
reprozip pack

reprounzip setup
reprounzip run

Come Try ReproZip!

Packing Experiments on Operating System *S*



System Call Tracing (reprozip trace)

- ReproZip transparently captures the provenance of the execution of the experiment, i.e., all the required information to correctly reproduce the experiment, including data files, programs, library dependencies, and OS information.
- The execution trace is stored in SQLite.

Provenance Analysis (reprozip trace)

- Given the files that were read and using the package manager of the OS, ReproZip identifies the software packages on which the experiment depends. ReproZip also uses some heuristics to identify input and output files.
- All the required information is written to a human-readable configuration file.

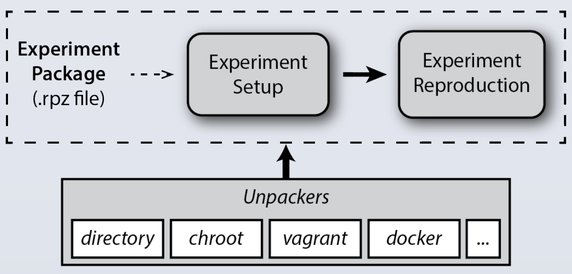
Package Customization

- The configuration file can be edited by researchers, e.g., to remove large files that can be obtained elsewhere, or to remove sensitive or proprietary information.

Package Generation (reprozip pack)

- All the required files are packed on the author's system *S* in a **.rpz** file.

Unpacking Experiments on Operating System *S'*



Unpackers

- *S* and *S'* are incompatible: *vagrant*, *docker*
- *S* and *S'* are compatible: *directory*, *chroot*, *vagrant*, *docker*

Experiment Setup (reprounzip setup)

- The experiment is automatically extracted and set up depending on the chosen unpacker.

Experiment Reproduction (reprounzip run)

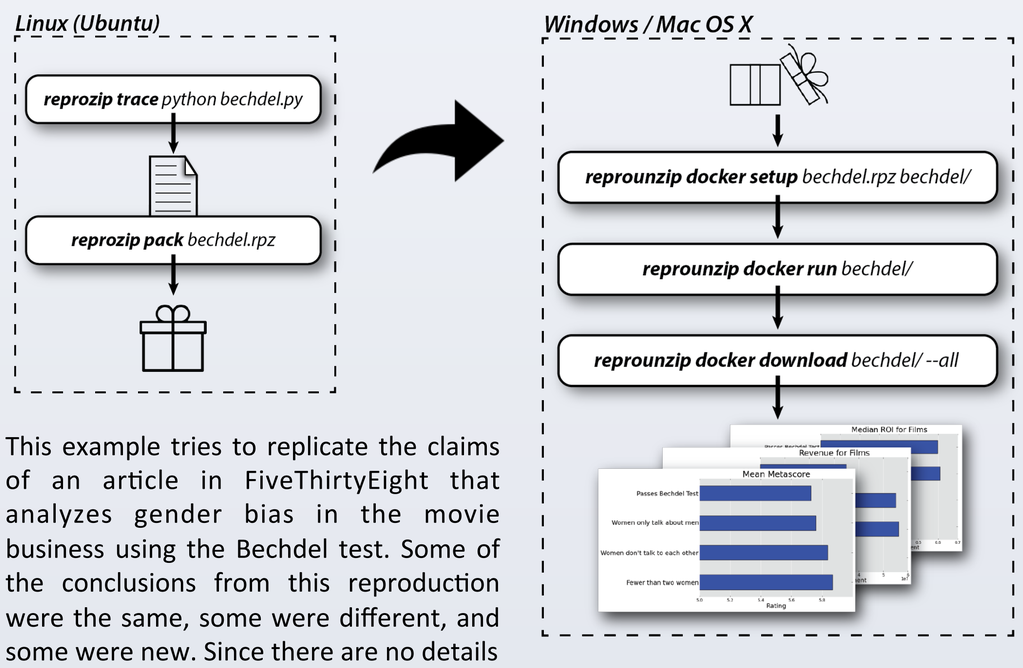
- The experiment is reproduced depending on the chosen unpacker. For instance, for *vagrant* and *docker*, this is done inside a virtual image and a Docker container, respectively, in an automatic and transparent way through *reprounzip* and its command-line interfaces.

File and Dataflow Management

- Input files can be replaced using ***reprounzip upload***.
- Output files can be retrieved using ***reprounzip download***.
- ReproZip can also derive a specification of the experiment for the VisTrails system, which represents the original workflow in a GUI and enables the dataflow to be modified to explore different techniques, perform analyses, and reuse some of the steps for your own research.

Example

This data journalism example is available at *http://bit.ly/BechdelFiveThirtyEight*



This example tries to replicate the claims of an article in FiveThirtyEight that analyzes gender bias in the movie business using the Bechdel test. Some of the conclusions from this reproduction were the same, some were different, and some were new. Since there are no details on the analysis performed in the original article, it is difficult to know why some of the conclusions differ.

Use Cases

- ReproZip supports a wide range of experiments, including client-server scenarios, experiments with databases, and graphical and interactive tools.
- ReproZip has been used by the Information Systems Journal to reproduce the results of published articles.
- ReproZip was recommended by the ACM SIGMOD 2015 Reproducibility Review.
- ReproZip has been listed on the Artifact Evaluation Process guidelines.

Acknowledgments

This work was supported in part by NSF awards CNS-1229185 and CI-EN-1405927, and by the Moore-Sloan Data Science Environment at NYU.

References

[1] ReproZip's Homepage: *https://vida-nyu.github.io/reprozip/*

[2] ReproZip Examples and Demos: *https://github.com/ViDA-NYU/reprozip-examples*

[3] F. Chirigati, R. Rampin, D. Shasha, and J. Freire, "ReproZip: Computational Reproducibility with Ease" In: Proceedings of SIGMOD'16, Demo Session, 2016.