

Reproducibility, Preservation, and Access to Research with ReproZip and ReproServer

Vicky Steeves^{1,2} | Librarian for Research Data Management & Reproducibility
Remi Rampin² | Research Software Engineer

¹Division of Libraries, ²Center for Data Science | New York University

Slides: osf.io/4uc3p

Reproducibility....



As all things, reproducibility is defined as a spectrum

Reviewable Research: Sufficient detail for peer review & assessment.

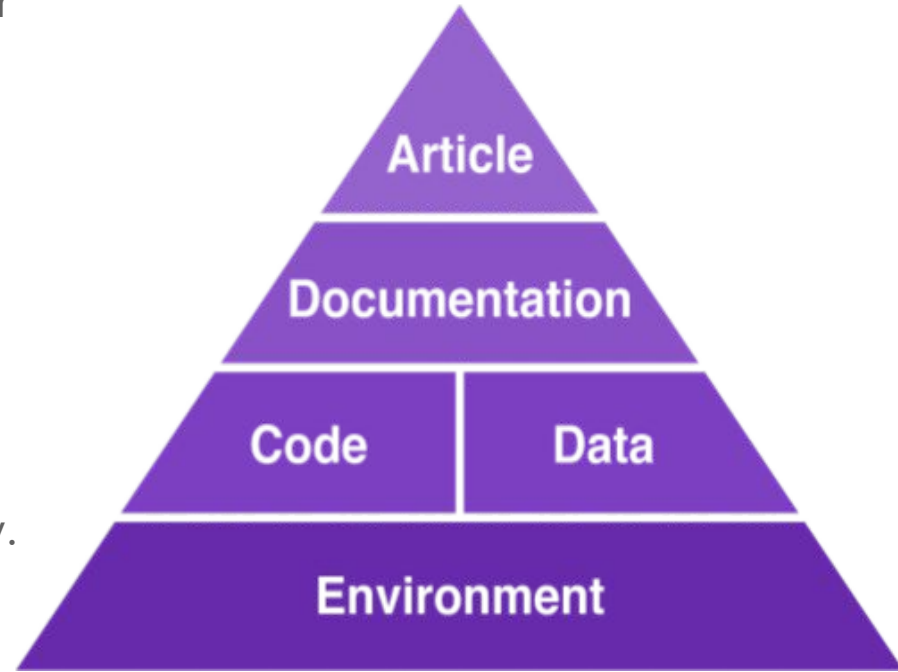
Replicable Research: Tools are available to duplicate the author's results using their data.

Confirmable Research: Main conclusions can be attained independently without author's software.

Auditable Research: Process & tools archived such that it can be defended later if necessary.

Open/Reproducible Research: Auditable research made openly available

[Stodden et al ICERM report \(2013\)](#)



ReproZip tries to solve...

Workload & Time Challenges

It is a time commitment to get data and code ready to share, and to share it

Otherwise known as...

the Incentive Problem

Reproducibility takes time, and is not always valued by the academic reward structure

“Insufficient time is the main reason why scientists do not make their data and experiment available and reproducible.”
Carol Tenopir, Beyond the PDF2 Conference

“77% claim that they do not have time to document and clean up the code.”
Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

ReproZip tries to solve...

Technical Obsolescence

Technology changes affect the reproducibility

Normative Dissonance¹

Espoused values don't always match practice

Otherwise known as...

The Pipeline Problem

Reproducibility requires skills that are often not included in most curriculums!

“It would require huge amount of effort to make our code work with the latest versions of these tools.” Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

¹<https://www.ncbi.nlm.nih.gov/pubmed/19385804>

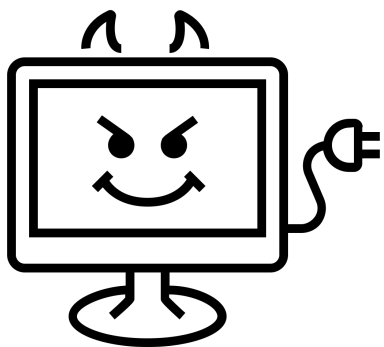
Even if runnable, results may differ...

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

We investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). **Significant differences** were revealed between **FreeSurfer version v5.0.0 and the two earlier versions**. [...] About a factor two smaller differences were detected between **Macintosh and Hewlett-Packard workstations** and between **OSX 10.5 and OSX 10.6**

The main problem ReproZip solves

Dependency Hell



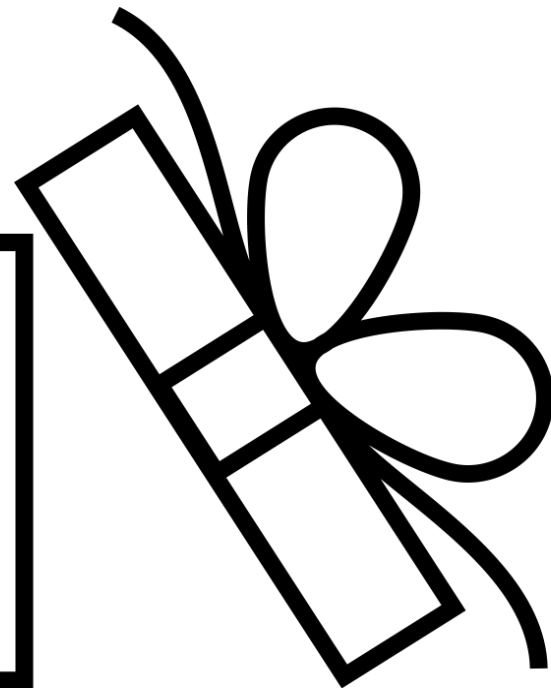
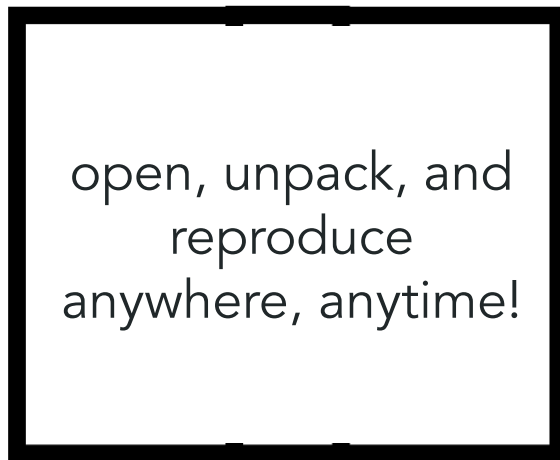
You cannot expect people to find all the chains of dependencies!

You cannot expect people to install all the dependencies and your code to run smoothly!



Gap: tools that can automatically capture all the dependencies in the original environment and automatically set them up in another environment

ReproZip *The Reproducibility Packer!*



ReproZip: Reproducibility in 2 Steps

Packing



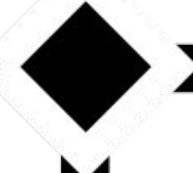
Linux



ReproZip Package



*data files, libraries,
environment variables, etc.
required to reproduce
the research*



Unpacking

Windows



Linux



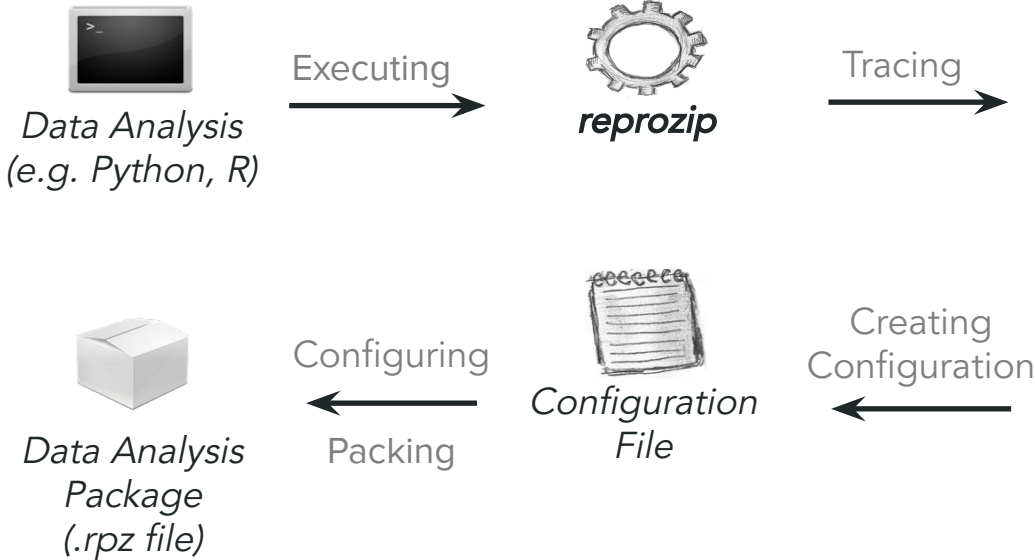
Mac OS X



*open, unpack, and
reproduce anywhere, anytime!*

Packing Your Work

Computational Environment **E** (Linux)



Experiment Provenance

Data

Input files, output files, parameters

Workflow

Executable programs and steps

Environment

Environment variables,
dependencies, ...

ReproZip can pack:

Data analysis scripts / software

(any language, you name it!)

Graphical tools

Interactive tools

Client-server applications (including databases)

Jupyter notebooks

MPI experiments (setting up the experiment can be involved but...)

... and much more!

Current Use Cases:

Academic Use Cases

- Recommended by the [Information Systems Journal](#), Reproducibility Section
- Recommended by the [ACM SIGMOD Reproducibility Review](#)
- Listed on the ACM [Artifact Evaluation Process Guidelines](#)

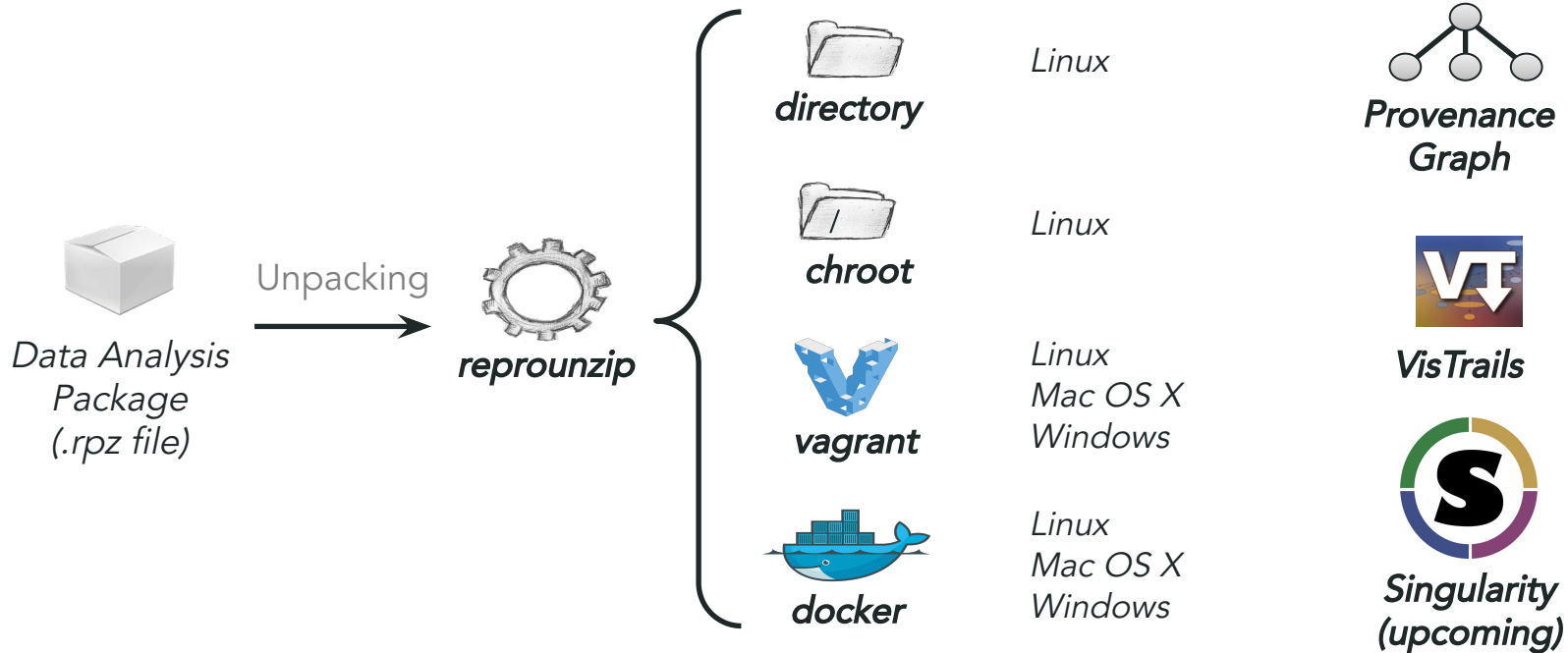
Outside Project Integration

- Integrated as a component of [CoRR](#)
- Archiving data journalism apps, e.g.: [Stacked Up](#)
- Used by [neurodocker](#) to build minimal Docker images

... and many more!

Unpacking Research

Potentially in a different environment / Operating System



Users can reproduce *and extend* the original work

Download Outputs



Upload New Inputs



Why we think our approach is good for preservation

Well-bundled:

- Captures **everything** your work touches, which is what it needs to rerun!
With lots of **extremely** technical metadata!

Generalizable:

- The RPZ format is simple but effective and very generalizable. It can interoperate, be read/accessed by, and run with a lot of software

Future-proofing:

- We can always add/remove unpackers to give users in the future full access to the bundle. As long as there are VMs, containers, or Linux, we can re-execute the bundle contents

BUT WAIT!

What if you don't want to download
ReproUnzip and Docker or Vagrant??

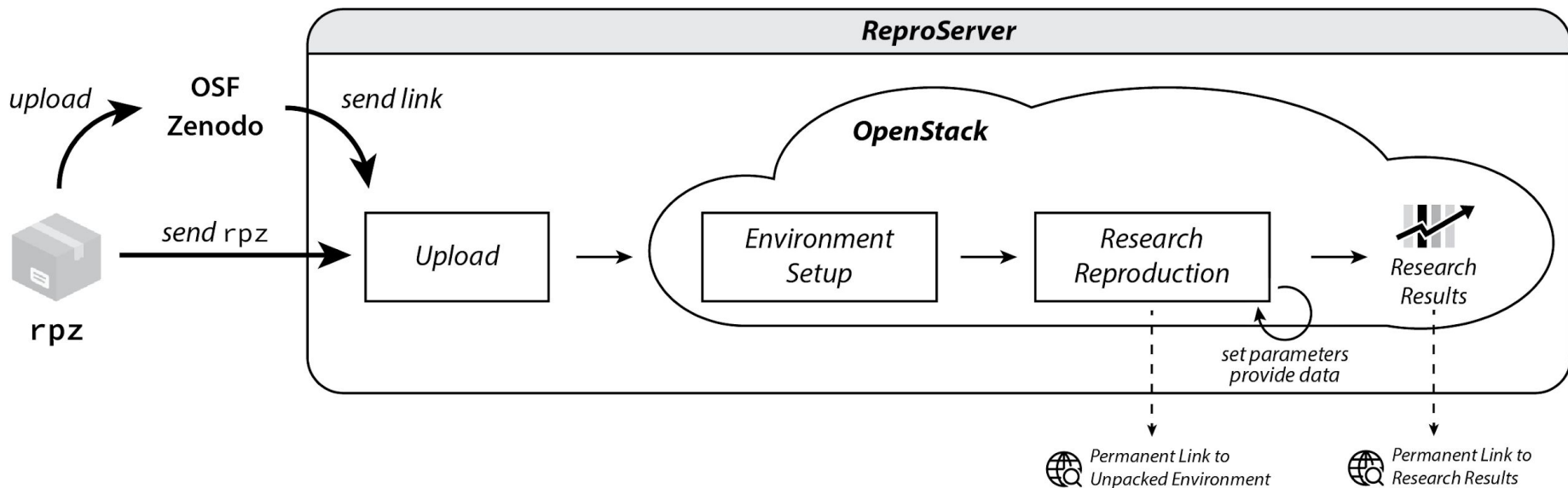
ReproServer to the rescue! No preservation w/o access!



ReproServer, reproducibility in-browser!



ReproZip + ReproServer = Access to Reproducible Work!





ReproServer -- simplifying access

- Runs ReproZip packages **in the browser**, no local software needed
- Allows **changing** input data & configs
- Gives you **a URL to include in papers** to reproduce your experiment
- Offloads archiving responsibility to people who are good at it (ayo)
- **No lock-in**: build on your laptop, pack automatically, reproduce anywhere

ReproServer Unpack

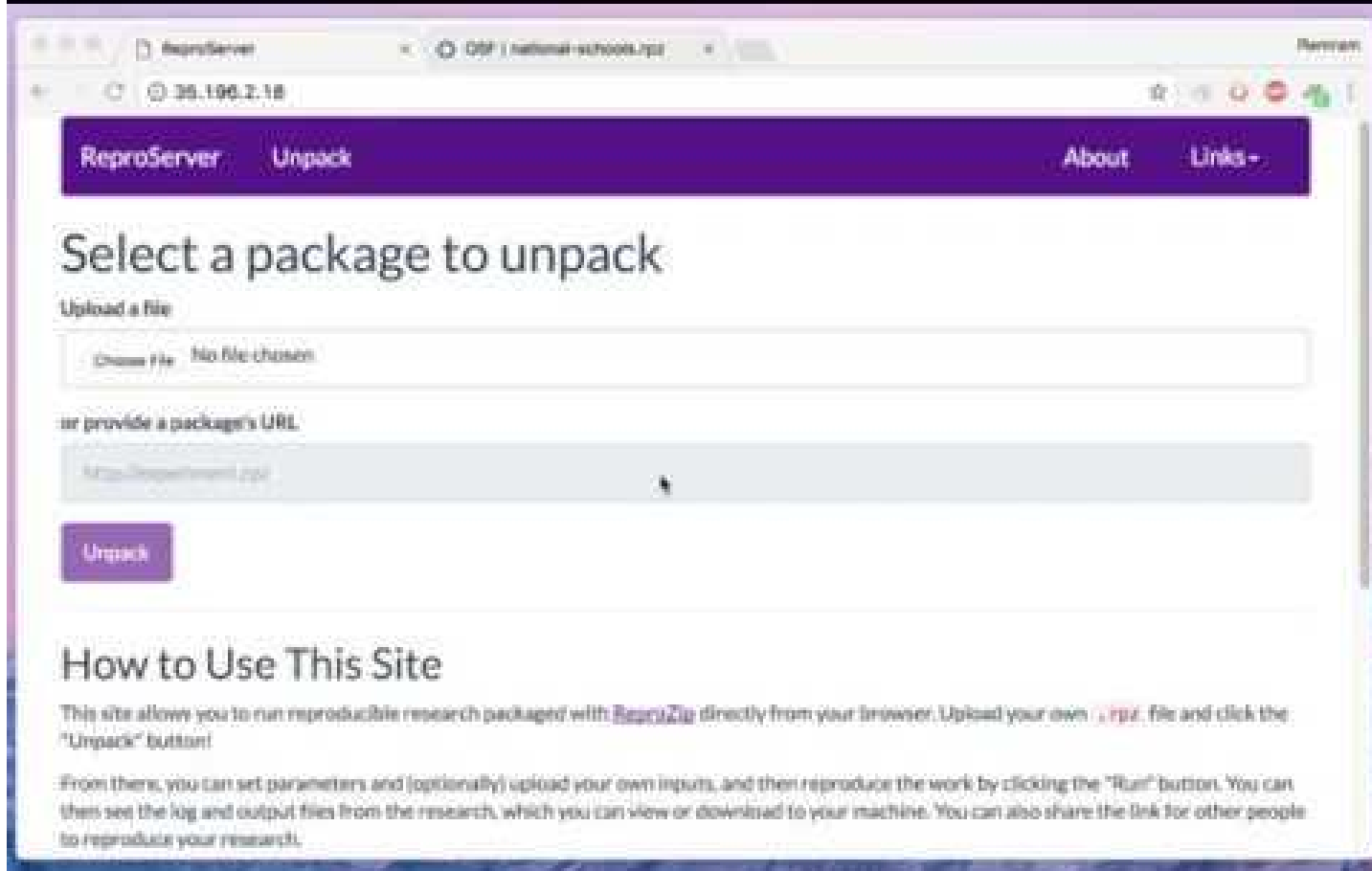
Select a package to unpack

Upload a file

No file chosen

or provide a package's URL

Unpacking local RPZ with ML Scripts from ReproServer



The screenshot shows a web browser window with the address bar displaying "35.190.2.18". The page title is "ReproServer" and the main heading is "Unpack". There are navigation links for "About" and "Links". The main content area is titled "Select a package to unpack" and contains two input fields: "Upload a file" (with a "Choose file" button and "No file chosen" text) and "or provide a package's URL" (with a text input field containing "https://reproserver.org"). Below these fields is a purple "Unpack" button. The bottom section is titled "How to Use This Site" and contains two paragraphs of text explaining the site's functionality.

ReproServer Unpack About Links -

Select a package to unpack

Upload a file

Choose file No file chosen

or provide a package's URL

https://reproserver.org

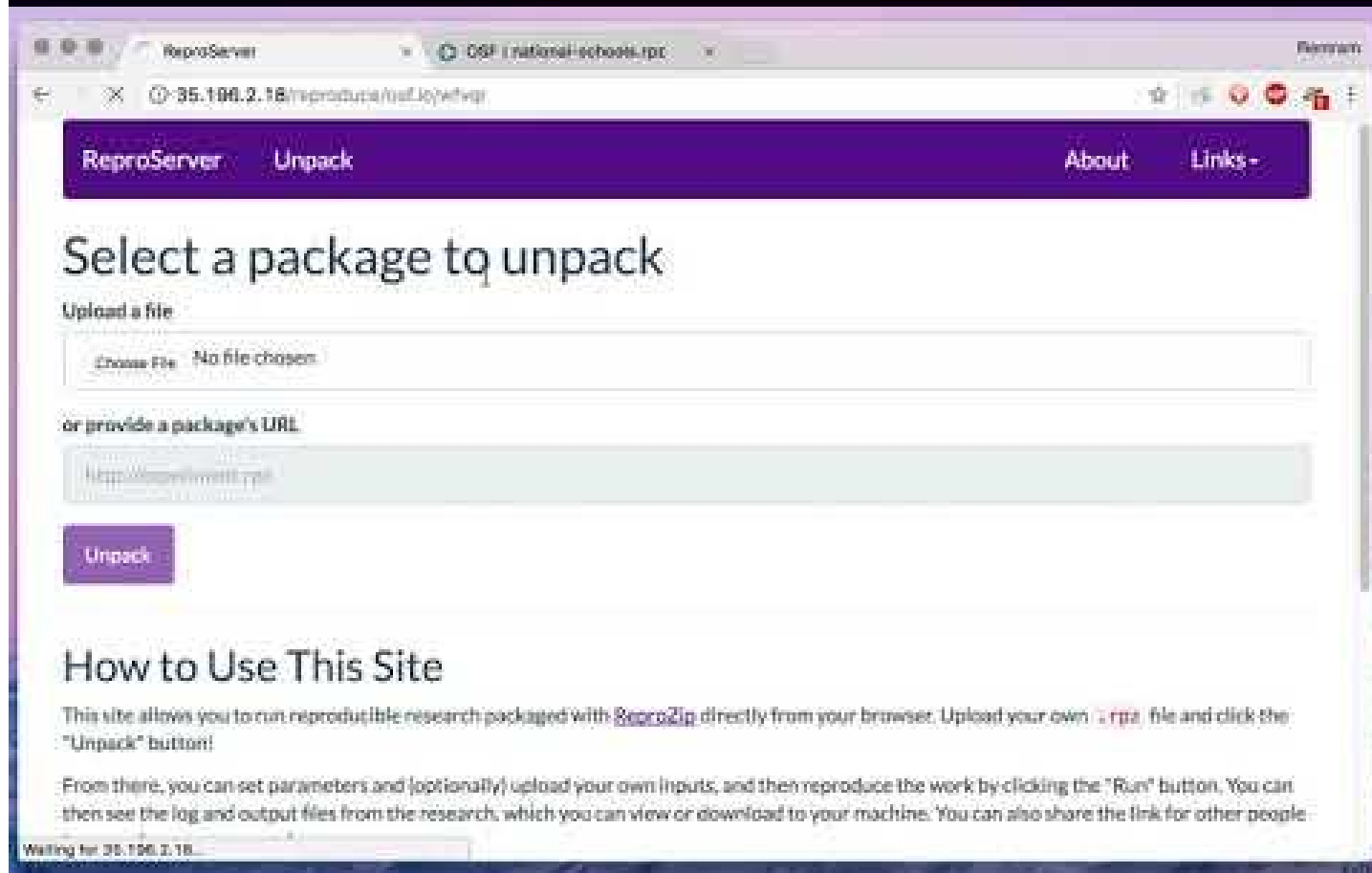
Unpack

How to Use This Site

This site allows you to run reproducible research packages with [ReproZip](#) directly from your browser. Upload your own `.rpz` file and click the "Unpack" button!

From there, you can set parameters and (optionally) upload your own inputs, and then reproduce the work by clicking the "Run" button. You can then see the log and output files from the research, which you can view or download to your machine. You can also share the link for other people to reproduce your research.

Unpacking R plots in RPZ bundle **DIRECT** From the OSF

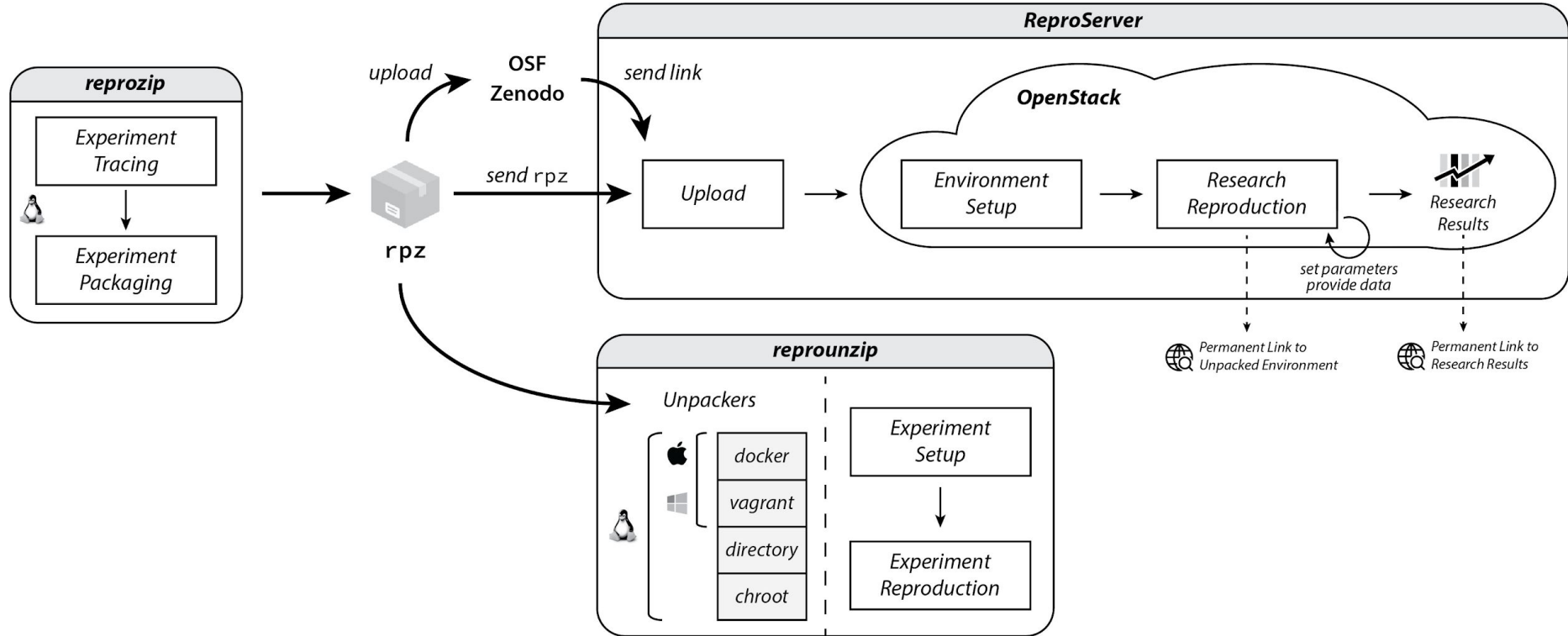


The screenshot shows a web browser window with the URL `35.196.2.18/reproducible/replj/efvgr`. The page has a purple header with the text "ReproServer" and "Unpack" on the left, and "About" and "Links -" on the right. The main heading is "Select a package to unpack". Below this, there are two options: "Upload a file" with a "Choose File" button and "No file chosen" text, and "or provide a package's URL" with a text input field containing `http://osf.io/mtt8yq/`. A purple "Unpack" button is positioned below the URL field. At the bottom of the page, there is a section titled "How to Use This Site" with two paragraphs of text. The first paragraph states: "This site allows you to run reproducible research packaged with `ReproZip` directly from your browser. Upload your own `.rpk` file and click the 'Unpack' button!". The second paragraph states: "From there, you can set parameters and (optionally) upload your own inputs, and then reproduce the work by clicking the 'Run' button. You can then see the log and output files from the research, which you can view or download to your machine. You can also share the link for other people". At the very bottom, there is a status bar that says "Waiting for 35.196.2.18".

ReproZip + ReproServer = Preservation + Access!

- **ReproZip** provides local, non-locked-in, reproducible packing of research -- easily integrated into existing workflows :)
- **ReproServer** provides a way for other users to interact with RPZ bundles from the comfort of their browser; easing access, review, and reuse of research materials
 - BONUS: see the work in the original computational environment
 - BONUS: can read in RPZ bundles from wherever they live (no duplicate upload necessary if in a repository)
 - BONUS: can be run locally at your institution (e.g. don't have to rely on centralized infrastructure not controlled by y'all)

The full ReproZip-ReproServer Ecosystem



Other Resources for ReproZip & ReproServer

ReproZip Website:

reprozip.org

ReproZip Examples:

examples.reprozip.org

ReproZip GitHub:

github.com/VIDA-NYU/reprozip

ReproServer GitHub:

github.com/VIDA-NYU/reproserver

ReproZip packing/unpacking:

goo.gl/o1Hqrx

Website packing: goo.gl/yMEOZJ

Jupyter notebook packing/replay:

goo.gl/NvMHnw

ReproServer demo:

goo.gl/Wk7Xnz

ReproServer OSF integration:

goo.gl/XfF78z

Summary:

- **ReproZip** provides the preservation + reproducible bundle of work from researchers
- **ReproServer** provides easier access to the materials of ReproZip bundles in-browser
- No preservation w/o access!

Thank You:

Fernando Chirigati, ReproZip OG dev & team member!

Juliana Freire, ReproZip PI

Moore & Sloan, for the green

Our users, for their feedback and continued help in dev!
