Packing Experiments for Sharing and Publication

Fernando Chirigati  
NYU-Poly  
fchirigati@nyu.edu

Dennis Shasha  
NYU  
shasha@cs.nyu.edu

Juliana Freire  
NYU-Poly  
 juliana.freire@nyu.edu

THE NEED FOR COMPUTATIONAL REPRODUCIBILITY

- The standard of having reproducible experiments, a long tradition in natural science, has not been adopted for computational experiments.
- Researchers often have to rely on figures, plots and tables presented in papers, which loosely describe the results. Consequently, these results are difficult to reproduce, leading to a credibility crisis in computational science [1].
- There are two main reasons why reproducibility is often not adopted:
  - Authors find it difficult to generate a compendium for their experiment
  - Reviewers and collaborators have difficulties trying to reproduce and verify the results, even when the compendium is available.

REPROZIP

ReproZip is a general tool for packing reproducible research that
- tracks operating system calls and creates a reproducible package from command-line executions, in the author’s environment E, with all the required files to run the experiment (packing step). Authors do not need to port their experiment to a specific tool.
- generates a workflow specification to help reviewers and collaborators explore and verify the results, facilitating the review process.
- extracts files and workflow on another environment E' (unpacking step).

Packing Experiments

Provenance Capture
- ReproZip transparently captures the provenance of the execution of the experiment.
- It uses SystemTap [2] to trace system calls and capture all the required information to correctly reproduce the experiment.
- The execution trace is stored in MongoDB [3], a NoSQL database

Provenance Analysis
- ReproZip extracts the trace data to create a provenance tree. Each node in the tree stores information about an OS process, such as file reads, file written and command-line arguments - if a process p spawns a process p’, an edge is inserted between their corresponding nodes.
- The provenance tree is traversed to identify programs, input files, output files and dependencies.

Package Generation
- The identified binaries and files are used to derive a workflow specification for the experiment.
- All the required files, together with the workflow, are packed on the author’s environment E using the original directory structure.

Unpacking Experiments

Package Extraction
- Given a reproducible package generated on environment E, ReproZip extracts all the files in a single directory on environment E', without interfering with this environment.
- The workflow is automatically configured to point to the correct files inside the unpacked directory.
- Paths defined in environment variables for the experiment are adjusted to use the experiment directory in E'.

REFERENCES