# Data Repository for Reproducible Research

Jerod Weinman
jerod@acm.org

October 21, 2014

**Abstract**

To promote reproducibility, since 2009 all my empirical research has been cached in a data repository that captures all source code and data dependencies automatically or semi-automatically. Our principle goals are to modularize, document, and preserve all aspects of experimental data.

So that others may benefit from or adapt my methods, I provide a bit of motivation and explanation here, along with the full description of the parameters and guidelines of the data repository (provided to my research students) and source code for the utilities.

## Overview

Peng and Eckel (2009) provide a framework for understanding reproducible research, as shown in Figure 1, at right. Our framework models theirs closely enough. Computations are cached in immutable objects called collections, and subsequent collections may utilize the results of other collections in a large dependency graph. In addition, version numbers of source controlled code are documented for reproducibility.

Our structure is overlaid on Peng and Eckel's graph overleaf, where we divide collections into a tripartite hierarchy:

`raw` Data external to our repository or manually generated and therefore not reproducible.

`proc` Data processed in a straightforward manner with a broad array of experimental uses.

`experiments` Data processed with a targeted test or particular experimental design.

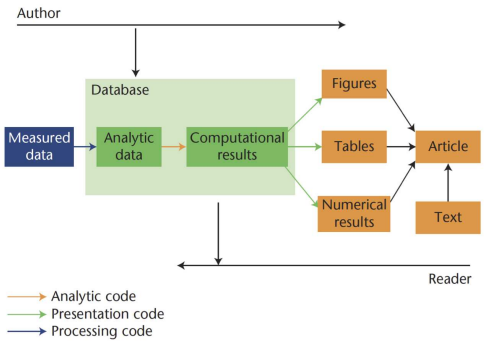


Figure 1: Reproducible research pipeline. From R.D. Peng & S. P. Eckel. (2009). *Comput. Sci. Eng. 11*(1), p. 28 . © IEEE.

## Documents and Code

The following example documents and code accompany this overview to provide additional detail and guidance.

**Data Repository Format and Management** Formal description and management practices.

**Reproducible Research** Discussion slides introducing the ideas to students.

**Data Repository** Repository skeleton, collection template, and processing utilities outlined in the document above.

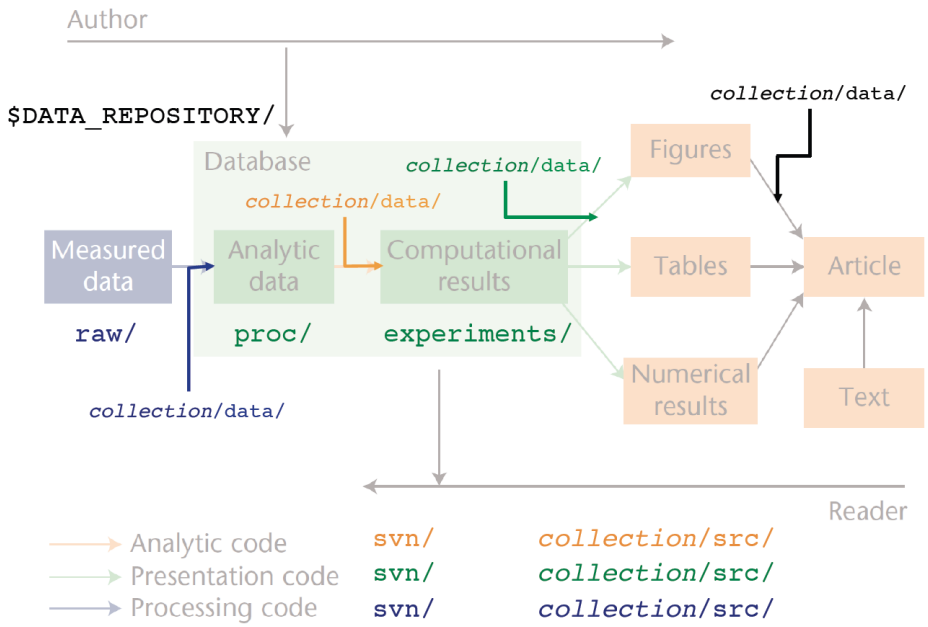


Figure 2: Data repository view of reproducible research pipeline (after Peng & Eckel).

## Example

The data repository (subset) included in the archive contains the raw data, processed counts, and final objects for a character bigram model of English text used in Weinman et al. (2014).

- `raw/text/books-20050601120000`
- `proc/text/bigrams-20090618132847`
- `experiments/text/meta/corpus_ten_fold_split-20090708080932`
- `experiments/text/ngrams/bigrams/tied_nums_intracase_L1_validation-20090708075734`

The astute reader will notice the time stamp on the bigram collection precedes the ten-fold-split. That is because the former called for the latter after some thought to generalization. One might also argue that the split should be in `proc`, and rightfully so.

## References

[1] Peng, R. D. and Eckel, S. P. (2009). Distributed Reproducible Research Using Cached Computations. *Computing in Science & Engineering, 11*(1), 28–34. doi:10.1109/MCSE.2009.6

[2] Weinman, J. J., Butler, Z., Knoll, D., and Feild, J. (2014). Toward Integrated Scene Text Reading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(2), 375–387, doi:10.1109/TPAMI.2013.126