Geophysical Research Abstracts
Vol. 14, EGU2012-11058, 2012
EGU General Assembly 2012

The Production Data Approach for Full Lifecycle Management

J. Schopf
Woods Hole Oceanographic Institution, Woods Hole, MA 02543 United States (jmschopf@gmail.com)

The Production Data Approach for Full Lifecycle Management

Dr. Jennifer M. Schopf

Woods Hole Oceanographic Institution

jmschopf@gmail.com

The amount of data generated by scientists is growing exponentially, and studies have shown [Koe04] that unarchived data sets have a resource half-life that is only a fraction of those resources that are electronically archived. Most groups still lack standard approaches and procedures for data management. Arguably, however, scientists know something about building software. A recent article in *Nature* [Mer10] stated that 45% of research scientists spend more time now developing software than they did 5 years ago, and 38% spent at least 1/5th of their time developing software. Fox argues [Fox10] that a simple release of data is not the correct approach to data curation. In addition, just as software is used in a wide variety of ways never initially envisioned by its developers, we're seeing this even to a greater extent with data sets.

In order to address the need for better data preservation and access, we propose that data sets should be managed in a similar fashion to building production quality software. These *production data sets* are not simply published once, but go through a cyclical process, including phases such as design, development, verification, deployment, support, analysis, and then development again, thereby supporting the full lifecycle of a data set.

The process involved in academically-produced software changes over time with respect to issues such as how much it is used outside the development group, but factors in aspects such as knowing who is using the code, enabling multiple developers to contribute to code development with common procedures, formal testing and release processes, developing documentation, and licensing. When we work with data, either as a collection source, as someone tagging data, or someone re-using it, many of the lessons learned in building production software are applicable. Table 1 shows a comparison of production software elements to production data elements.

Table 1: Comparison of production software and production data.

Production Software	Production Data
End-user considerations	End-user considerations
Multiple Coders: Repository with check-in procedures Coding standards	Multiple producers/collectors Local archive with check-in procedure Metadata Standards
Formal testing	Formal testing
Bug tracking and fixes	Bug tracking and fixes, QA/QC
Documentation	Documentation
Formal Release Process	Formal release process to external archive
License	Citation/usage statement

The full presentation of this abstract will include a detailed discussion of these issues so that researchers can produce usable and accessible data sets as a first step toward reproducible science. By creating production-quality data sets, we extend the potential of our data, both in terms of usability and usefulness to ourselves and other

researchers. The more we treat data with formal processes and release cycles, the more relevant and useful it can be to the scientific community.

References

[Fox10] Fox, P., "Why the term 'data publication'?", http://tw.rpi.edu/weblog/2010/12/14/why-the-term-data-publication/

[Koe04] Koehler W., "A longitudinal study of Web pages continued: a consideration of document persistence", *Information Research*, 9 (2), January 2004. Also available at http://informationr.net/ir/9-2/paper174.html

[Mer10] Merali , Z., "Computational science: ...Error: . . . why scientific programming does not compute", *Nature* 467, 775-777 (2010). doi:10.1038/467775a . Also available at http://www.nature.com/news/2010/101013/full/467775a.html