Ten Habits of Highly Effective Data

Anita de Waard

Elsevier Research Data Services, Jericho, VT 05465, USA
a.dewaard@elsevier.com

One of the main goals of the current trend to store and share outputs of experimental research (cf. the OSTP memo, the NIH B2DK project, and the NSF Research Data sharing policy) is to encourage the reuse of that data. Therefore, it is important that usage is taken into account when designing systems that store and create data. Conversely, the many parties interested in data science [Press, 2013] should care about how, and that that research data gets stored in a way to be optimally usable downstream. Alignment of the ten aspects listed below could support optimal data reuse and lead to the development of better systems. There is an intended hierarchy to these aspects, akin to the Maslow hierarchy of human needs: each builds on and requires the aspects preceding them.

1. Preserved: existing is some format

   At this moment, the majority of data in labs all around the world is not saved. The Data Rescue Award showcased several attempts in the Earth Sciences to rescue 'dark data'.

2. Archived: existing in a long-term durable format

   Data needs to be preserved in a format-independent manner or risk being lost forever. The Olive Executable Archive brings a number of arcane operating systems back to life as Virtual Machines.

3. Accessible: available to others than the researcher

   The Toronto workshop proposed prepublication sharing for genomics and proteomics data. Data sharing systems such as Labfolder allow the creation of private groups that can be expanded to public spaces, minimizing the

4. Comprehensible: understandable to others

   As an example to enable comprehension, the Urban Legend Data Dashboard aims to make data directly available for analysis to researchers through a web-based visualization.

5. Discoverable: can be indexed by a search engine

   A key issue will be how data can be found that is not directly linked to a publication. Recent funding proposals encourage the development of such data search engines; initiatives such as the National Data Service aims to provide a data 'discovery layer'.

6. Reproducible: allows others to reproduce the experiment

   The Reproducibility Initiative enables the reproduction of experiments for a fee. Activities such as the Force11 Resource Identification Initiative aim to identify key experimental components to improve reproducibility.

7. Trusted: validated by some authority, provenance known

   A key issue will be when scientists trust data created by others; adoption seems to be domain-specific. A key question is whether data curation can be made much easier – and cheaper – with tools such as Data Tamer [Stonebraker et al., 2013].

8. Citable: able to link to dataset and track citation

   The Force11 Data Citation Principles compile existing guidelines for data citations; work on widespead implementation is being undertaken with a large group of publishers.

9. Usable: allow tools to run over the data

   The big question is how all this disparately collected data will be used to make new and better science. Possibly, new advances in IR and simple metadata standards such as http://schema.org, which have enabled the explosion of community-driven creativity on the web might harness similar revolutions in science.

10. Integrated: upstream and downstream align

    A key feature of a successful data management systems would be that there is successful integration between the nine aspects described above.

In summary, we can improve the way that research data is created, stored, and managed, if systems for doing so are designed with usage into account. Similarly, data scientists should be more closely involved, and concerned with, in the details of data creation and storage. Workshops such as this one support communication between these different groups.

Reference

Gil Press, A Very Short History Of Data Science, Forbes Tech, 5/28/2013: http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

M. Stonebraker, D. Bruckner, et al., 'Data Curation at Scale: The Data Tamer System', 6th Biennial Conference on Innovative Data Systems Research (CIDR '13), Jan 6-9, 2013, Asilomar, CA USA https://cs.uwaterloo.ca/~ilyas/papers/StonebrakerCIDR2013.pdf