Looking Out for the Little Guy: Small Data Curation

by Katherine Goold Akers

EDITOR'S SUMMARY

While big data and its management are in the spotlight, a vast number of important research projects generate relatively small amounts of data that are nonetheless valuable yet rarely preserved. Such studies are often focused precursors to follow-up work and generate less noisy data than grand scale projects. Yet smaller quantity does not equate to simpler management. Data from smaller studies may be captured in a variety of file formats with no standard approach to documentation, metadata or preparation for archiving or reuse, making its curation even more challenging than for big data. As the information managers most likely to encounter small datasets, academic librarians should cooperate to develop workable strategies to document, organize, preserve and disseminate local small datasets so that valuable scholarly information can be discovered and shared.

KEYWORDS

data curation

research data sets

knowledge management

digital repositories

archives

academic libraries

It is perhaps no coincidence that academic librarians are accelerating their efforts to curate research data at the same time that attention is increasingly being focused on big data. But in a college or university setting, a preoccupation with big data may be unrealistic and unproductive. Instead, our concerns should be with small data and the challenges of managing a myriad of diverse and undocumented, yet small, datasets.

With big data at the forefront of discussion, it is easy to imagine that all scientists are routinely generating massive quantities of research data. And some are. But many of these big data producers work in federal or private labs that are out of the reach of academic librarians. Or they work in universities but are supported by grants enabling them to hire staff and build infrastructure necessary to manage their own data without assistance from librarians. What remains is the majority of researchers based in colleges and universities, who conduct several small, interrelated projects simultaneously, many of which are either not federally funded or not funded at all. Some of these projects culminate in journal articles and conference presentations, and others quietly phase out, but in either case, the data are rarely preserved or disseminated. Although science is rapidly evolving with advances in thinking and technology, it has not undergone a paradigm shift, at least not yet. Despite having more sophisticated tools, most scientists continue to make progress via incremental steps – series of observational, experimental or computational studies that each generate a relatively small amount of data. The same is true for researchers in the social sciences and humanities, who also typically produce modestly sized datasets or collections of digital objects. It is these small datasets that academic librarians are most likely to encounter [1].

It is also easy to get the impression that, as a general rule, big data are more important than small data. But the value of a dataset is not a function of its size. Certainly the curation of big data is vital in many cases. Consider the unique conglomeration of weather conditions leading up to the recent superstorm Sandy. Meteorological data captured from this event are not reproducible. If the data are not preserved, scientists lose opportunities to gain additional insights into how such storms arise and how to improve our weather prediction systems. However, big data often accumulate through unfocused sweeps of the environment at

Katherine Goold Akers has a background in psychology and neuroscience research. She is now a Council on Library and Information Resources (CLIR) postdoctoral fellow and e-Science librarian at Emory University. She can be reached at katherine.g.akers<at>emory.edu.

microscopic to astronomical scales, with potentially valuable information hidden in huge amounts of noise. Small data, in contrast, are often the result of highly focused, carefully designed studies with a limited number of observations. Not all small studies yield data that are highly valuable or worthy of long-term preservation, but they do tend to be the types of studies that explore new areas of inquiry or that narrow in on answers to long-standing questions. If preserved and meaningfully integrated, small data can be just as important as big data, if not more so.

Worry over the management of big data is justified, as it presents significant challenges in terms of storage, accessibility and analysis. But because big data are difficult to curate does not mean that the curation of small data is easy. In fact, in some situations, small data can be even more difficult to manage than big data. The instruments that churn out big data often produce files that are highly standardized and automatically accompanied by metadata. These files often fit nicely into established disciplinary repositories, allowing data to be preserved in perpetuity, linked to related data and accessed by the individuals who are best suited to re-use the data in new and meaningful ways. Small data is messier [2]. Any individual small research project can generate a surprising number of files in a variety of formats. Because the goal of academic researchers is to publish manuscripts as quickly as possible, they spend little if any time preparing their data for archival or re-use, meaning that most datasets lack adequate documentation or metadata. Some research areas have no appropriate disciplinary repositories. Whether the orphaned data make it into institutional repositories, it is unclear how these detached, diverse datasets can be synthesized into something that is greater than the sum of its parts.

Because academic librarians are more likely to encounter small data in their colleges and universities, we should focus on developing approaches to the documentation, organization, preservation and dissemination of small datasets that have no permanent home outside of the labs and offices in which they were born. The effective and efficient curation of small data may prove to be an immense feat, but one that offers profound opportunities for librarians to meet an unfulfilled need and to forge new paths in the discovery and sharing of scholarly information. ■

Resources Mentioned in the Article

[1] Salo, D. (2010). Retooling libraries for the data challenge. *Ariadne, 64*. Retrieved January 3, 2013, from www.ariadne.ac.uk/issue64/salo.

[2] Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280-299.