Data Management as Bibliography

by Michael Buckland

EDITOR'S SUMMARY

A critical element in the products of research projects is the data set, largely treated as a marginal appendage to the written record. With the requirement for data management plans in grant proposals to the National Science Foundation and National Institutes of Health, the issue of managing research data sets is gaining attention. Sharing primary data enables a greater return on the original investment, expanded discovery and fertilization of ideas across disciplines. The major impediment is lack of an infrastructure to archive non-textual data sets, in addition to hurdles including locating data, deterioration, format standardization, permission for use and suitability. It is essential that institutional policies recognize data sets as a key product of intellectual work worthy of inclusion in a bibliography. This must be followed by action to promote their preservation, metadata enrichment, cross-lingual interoperability, geographic and temporal coding and data provenance and to contend with changes over time.

KEYWORDS

research data sets
data set management
information reuse
library and archival services
organization of information

Michael Buckland is emeritus professor at the School of Information and co-director of the Electronic Cultural Atlas Initiative, University of California, Berkeley, CA 94720-4600. He can be reached at buckland<at>ischool.berkeley.edu.

Research projects typically generate data sets, but in practice it is commonly impractical for anyone else to attempt to re-use these data. The National Science Foundation and the National Institutes of Health now mandate that grant proposals include a data management plan explaining how data sets generated as part of the project would be preserved and made accessible [1, 2, p. II-19]. The requirements are not very specific, and it is not clear that there is much accountability. Nevertheless, these mandates are a major strategic move to address an issue that has finally begun to attract attention with high level reports and numerous conferences. In 2009 the National Academies of Science launched a new Board on Research Data and Information.

Science and engineering are constructive enterprises evolving through hypotheses and model building, trial and error, testing and revision. For this reason shared access to the record of prior work is critical. Historically, the record has been primarily textual in the form of published technical reports, articles, conference papers, books and other genres, although there were always some non-textual records, such as collected specimens.

Print-on-paper materials are made accessible through a slowly evolved infrastructure of scholarly norms (for example, acknowledgement and citation), genres of technical writing, specialized publishers and distribution channels, libraries, and bibliographies, catalogs and indexes. The infrastructure for publishing and bibliographical access was established by scholars, societies, librarians and publishers. During the second half of the 20th century digital methods made new techniques feasible. (One thinks of *Chemical Abstracts*, *Medline* and the *Science Citation Index*.) It is a creaky system but it works.

No comparable infrastructure is in place yet for data sets, which undermines the credibility of even well-intentioned data management plans.

It is not known how bad the situation really is. If one were to pick a random selection of papers reporting the results of federally funded projects completed five or 10 years ago and sought to re-use the data sets they were based on, the effort would probably generate more frustration and embarrassment than success.

If these data sets are regarded as illustrative appendages to a definitive textual record then this situation is regrettable. But the situation is worse than that because the practice of science and engineering has also been transformed by the pervasive adoption of digital computation and communication. The potentially useful record of science is increasingly not the written reports but (mainly non-textual) digital data sets of many kinds: the raw material, the operations upon it and progressively more refined derivations can be beneficially shared and built upon by other researchers, not only in the same field but also in adjacent fields. This potential extends the impact and broadens the evidence in ways not practical using textual reports alone and has enabled a steep rise in computationally intensive, data-centric science. The potential now exists, therefore, for a far greater return on investment in research, but there is a requirement: the infrastructure of well-developed work practices, publication norms, libraries and bibliographical access that evolved to create and sustain an accessible archive of the *literature* of each field has to be complemented by a corresponding set of work practices and infrastructure for the archive of *non-textual digital data sets* that constitute an ever-increasing proportion of the record.

Researchers tend to work within domains and in relatively narrow research fronts with informal, interpersonal interaction within each specialty. Researchers know each other or, at least, of each other. They graduate from similar programs, work in teams, meet in conferences, read the same journals and correspond by email. These informal social networks strongly complement the formal channels of communication and documentation. In interaction between research fronts, however, this informal social network is largely absent. Without membership in the same "invisible college" researchers are unlikely to know what they could ask for or whom they could ask. And they are less likely to receive cooperation.

Rich results can be obtained when researchers explore at or over the boundaries of their fields and encounter ideas and/or data that are relevant but new and different (for them). This potential reward is why research funders and academic planners have long tried to induce more interdisciplinary interactions in the resolutely discipline-based academic environment. A resource that can be made to benefit more than one group yields a greater return on the investment.

There are examples of good practices in the very largest of science projects and in social science numeric data series, but widespread and largely undocumented deficiencies elsewhere. The significance of the problem can be sensed by imagining that a large proportion of the textual record was written but never published and remains largely inaccessible or unintelligible. What a waste!

Identifying Impediments to Data Reuse

The use of data sets generated by others in the past can be impeded in many different ways – the hard drive crashed, and there was no back-up; the person who could give permission cannot be found and so on. There are some clearly distinct barriers to be overcome. Here is one typology:

1. Discovery: Does a suitable data set exist?
2. Location: Where is a copy?
3. Deterioration: Is the copy too deteriorated and/or obsolete to be usable?
4. Permission: May it be used?
5. Interoperability: Is it standardized enough to be usable with acceptable effort?
6. Description: It is clear enough what the data represent?
7. Trust: Are the lineage, version, and error rate understood and acceptable?
8. Use: Should I use it for my purpose?

In practice the answers are unlikely to be a simple yes/no. A positive answer is not, in itself, enough. Any significant effort required to achieve a positive outcome inhibits action. It is always situational: the willingness to invest effort depends on the perceived benefits of success and the known alternatives as well as the cost and the resources available. One may satisfice:

a less perfect result that requires less effort will often be a reasonable preferred course.

These questions form a chain: If you learn that a data set exists you may not be able to locate a copy; if you can locate a copy it may not be usable; if it is usable, you may not be able to obtain permission; and so on. Any problem might prevent re-use, but, if resolved, another might still prevent it.

These impediments are different in kind and require different kinds of solutions: policies, work practices, infrastructure and so on. For example, one repository accepted data sets with the condition that the permission of the depositing principal investigator (PI) was required for third-party use, but with no contingency policy for when the PI died or was unavailable. Some remedies are more feasible and/or more affordable than others.

A particular problem is that descriptive metadata sufficient for the original compiler of the data is likely to be insufficient for someone else, years later, who may not know what the compiler took for granted and left implicit rather than explicit [3].

Mapping Responsibilities

The final question – Should I use it for my purpose? – is different from the other seven because in this case responsibility rests with the potential user, the domain specialist him or herself. Yet the decision is influenced by the answers to the other seven questions for each of which there are identifiable specialists and institutional loci specializing, more or less, in providing support. These are more clearly identifiable for the textual record than for data sets. Traditionally bibliographies identify what resources exist, catalogs list where copies can be found and now search engines support both tasks. Publishers provide copies in the short term; libraries provide long-term access and so on. Identifying the corresponding actual and potential institutional loci for the provision of sustained access to data sets would be a practical approach. This identification would help identify the specialists and would also provide a framework for evolving the institutional infrastructure.

Defining *Bibliography*

In ancient Greece, a bibliographer was a "book-writer," a copyist who transcribed an existing book to make a new copy. When the word *bibliographer* came into use in Europe, it was used more or less interchangeably with librarian until Martin Schrettinger, Melvil Dewey and others developed library science as a distinctive technical field [4]. By the mid-20th century *bibliographical access* or simply, *bibliography* (when used in a broad sense) were terms of choice in the print on paper world for, loosely, the issues associated with the eight questions listed above. This usage is reflected in the subtitle of Patrick Wilson's classic analysis in 1968 of the problems of organizing and selecting documents: *Two Kinds of Power: An Essay on Bibliographical Control* [5]. But then terminology changed and this broad sense of "bibliography" was largely displaced by "organization of information" and similar phrases. By default, the term *bibliography* was increasingly associated with a narrower sense: the detailed examination of printed books as physical objects also known as "analytical bibliography" or "historical bibliography."

An eloquent protest against this narrow view can be found in D. F. McKenzie's 1985 Panizzi Lectures entitled *Bibliography and the sociology of texts* [6]. McKenzie, a specialist in historical bibliography and textual criticism, argues persuasively for a broader approach in two ways. First, bibliography should extend beyond the book itself to include its interpretation and social context. This expansion has happened. Second, *text* should be interpreted widely to extend beyond writing in the printed books to include other media, notably films, maps and digital data sets. On this front far more needs to be done.

Areas Needing Attention

Numerous areas, including the following, need attention in addition to the central issue of preservation of digital data:

- *Metadata enrichment.* How could existing metadata for re-usable data sets be improved or extended, cost-effectively, with clear separation and ancestry of both new and old and with maximal interoperability, using annotation techniques, namespaces and other Semantic Web elements?
- *Cross-lingual interoperability.* Strong cross-lingual issues arise when the metadata of two sets are in different languages such as English and German. But also, since language evolves within fields of discourse, a

weak linguistic mismatch occurs between the specialized terminologies in different specialties. Retrieval performance is sensitive to these dialect differences. Computational linguistics can help.

- *The economics of harmonization.* Standards constrain flexibility but achieve long-term economies and resource sharing through interoperability. Multiple trade-offs are involved.
- *Coherence*. When moving beyond text on paper, resources are less visible. There will need to more focus on issues and features common to most science data sets, notably
  - Where: place and spatial location, geo-referencing
  - When: periods and calendar time, geo-temporal encoding
  - Data provenance: the need to be able to trace data back to its origin and justification
  - Boundary issues through time: shifting political boundaries, unstable biological taxa, and so forth
  - Ontologies (controlled vocabularies) shared or interoperable across domains.

These issues apply also to textual resources.

Conclusion

The problem of access to the non-textual record is finally getting serious attention. The National Academy of Sciences has established a Board on Research Data and Information. Conferences on the topic have become frequent. NSF launched its Sustainable Digital Data Preservation and Access Network Partners (DataNet) program. Universities are developing data repositories and, significantly, funding agencies (notably NSF and NIH) are now *requiring* data management plans in all proposals. The requirements are still general and vague, but the mandate for significant change is clear. If these requirements were enforced and audited, it would provide an excellent driver for the development of the requisite infrastructure. In the meanwhile not only is the production of data sets increasing in scale, but so much more can be done with them – replicating experiments; obtaining better, broader evidence; subjecting them to visual analysis; indulging in computationally intensive research and so on.

Change will require both the development of new, improved, attractive practices and a process of facilitated adoption. It is high time to bring bibliography up-to-date to address the management of media, notably data sets. As McKenzie put it [p. 52], "Further neglect of them is inexcusable." And that was in 1985!

Acknowledgments

I am grateful for the help of Jeanette Zerneke and Evan M. Smith and for the support of the Coleman Fung and the Coleman Fung Foundation "Knowledge Unix" gift to the Electronic Cultural Atlas Foundation. ■

Resources Mentioned in the Article

[1] Office of Extramural Research, National Institutes of Health. (2007). *NIH data sharing policy.* Bethesda, MD: The Institutes. Retrieved June 19, 2011, from http://grants.nih.gov/grants/policy/data_sharing/index.htm.

[2] National Science Foundation. (January 2011). *Grant Proposal Guide* [Section IIC.2.j.]. Washington, DC: The Foundation. Retrieved June 19, 2011, from www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpgprint.pdf.

[3] Bowker, G. (2008). *Memory practices in the sciences.* Cambridge, MA: MIT Press.

[4] Blum, R. (1980). *Bibliographia: An inquiry into its definition and designations.* Chicago: American Library Association.

[5] Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control.* University of California Press. Full text available at http://books.google.com.

[6] Mckenzie, D. F. (1986). *Bibliography and the sociology of texts.* London: The British Library.