Knowledge Rules: Curating Knowledge in the Social Sciences
Social Sciences Research Council Meeting, 2 May 2016, New York Public Library

Not Fade Away: Social Science Research Data in the Digital Era

Christine L. Borgman, UCLA

Not Fade Away[1] (Hardin & Petty, 1957)

I wanna tell you how it's gonna be
You're gonna give your love to me
I wanna love you night and day
You know my love will not fade away
You know my love will not fade away
Not fade away
My love's bigger than a Cadillac
I try to show you but you drive me back
Your love for me has got to be real
You're gonna know just how I feel
Love's real, not fade away
Not fade away

Introduction

Social scientists face several competing challenges for their research data. One is the pressure to make their data open in response to mandates from funding agencies, journals, and science policy makers. Second is the lack of resources – human, technical, economic, and institutional – to make their data open. Third is that good reasons exist to maintain control of their data, whether to protect the confidentiality of human subjects, to gain competitive advantage over other researchers, or the sheer difficulty of extracting data from the contexts in which they originated.

Competing scenarios

These are several competing scenarios for how those challenges may play out over the next decade or so.

Policy maker's ideal scenario

Social scientists will design their research studies to optimize the production of reusable data. Data will be released on a regular basis, no later than the time of publications or the

[1] Hardin, Charles and Petty, Norman, "Not Fade Away". 1957. The song was first performed by The Crickets and later covered by the Rolling Stones, Grateful Dead, Sheryl Crow, and many more. Rights are now held by Paul McCartney (Wikipedia, 2016).

end of grant periods. Data will be contributed to curated archives that will sustain them indefinitely, keeping them scientifically useful and available. Other social scientists, educators, government, business, and the public at large will reuse those data to create new knowledge and innovations. Authors will provide full citations to the data they produce and to they data they use, increasing traceability and discoverability.

Data librarian's ideal scenario

Social scientists will design their research studies to optimize the production of reusable data. They will produce their data using standard protocols, data structures, and non-proprietary software. Data sets will be documented thoroughly, including metadata for each variable, data cleaning procedures, handling of missing data, and transformations. All scripts and other algorithms used to analyze data will be provided. Codebooks will provide adequate description of the research design to make the study reproducible by others at least a decade later. All research memos, versions of papers and publications, and artifacts of the research life cycle will be provided. Each artifact will be assigned a permanent and unique identifier. Relationships between artifacts will be documented so that the graph of artifacts can be published. Researchers will submit their full portfolios of data to the archive in a timely manner, in a form that can be ingested following standard practices such as the Reference Model for an Open Archival Information System (Consultative Committee for Space Data Systems, 2012).

Social science researcher's ideal scenario

Social scientists will design their research studies to optimize innovation in data sources, theories, and methods. Research methods will be adaptable to context and conditions, while maintaining professional standards for reliability, validity, and protection of human subjects. Obtrusive methods, such as interviews, surveys, and ethnographies, will be used where appropriate. Similarly, unobtrusive methods such as gathering records of human behavior, past or present, in any medium in which recorded, will be used where appropriate. Researchers will acquire digital traces of human activity from whatever sensors or other devices collect them. Data will be aggregated and integrated from disparate sources. Novel instruments, protocols, and software tools will be employed to address new research questions. Researchers will hold intellectual property rights in their data and in their personalized methods. They will release data to peer reviewers or to individual requestors, but no sooner than at the time papers are submitted for publication. Licensing and human subjects protection procedures will apply. Potential reusers of data will be responsible for acquiring software and other tools necessary to use the data, and for all interpretations thereof.

Comparing scenarios

Needless to say, these three scenarios are fundamentally incompatible. Yet more incompatible scenarios can be generated from the perspectives of funding agencies, university administrators, private business, students, and other stakeholders. Specifics of the scenarios will vary greatly by research specialty, institution, career stage, funding source, and other factors.

These incompatibilities arise from differences in fundamental assumptions about data – assumptions that all too rarely are made explicit. The most problematic assumption is that "data" is an agreed concept. Both the policy maker and data librarian scenarios presume that data are bounded research products. Rather, almost anything can become data in the social sciences. One person's signal is another's noise. The researcher's scenario is based on the latter assumption. Identifying a new source of data, whether an obscure communication signal or a trove of archival documents, is itself a scholarly act.

A related conflict is between the value of standards in collecting and documenting data. The first two scenarios presume that researchers should strive for standards and consistency, on the grounds that systematic approaches increase reliability, integration, and reproducibility. The third scenario presumes that such standards severely limit the options for research design and hamper innovation.

Another conflicting assumption in these scenarios is whether data have meaning outside the contexts in which they originated. To the extent that data are research products that can be exchanged, then the open access expectations of the first two scenarios can be accomplished. To the extent that data reflect contextual understanding of a social situation, they will have little value for exchange or reuse. The researcher's scenario cuts both ways. By retaining control over the reuse of data, the researcher is asserting the importance of context. By aggregating data from multiple sources, the researcher is treating data as exchangeable products.

Lastly is the difference in assumptions about whether research data are public or private goods. Until recently, most fields viewed research data as private goods, part of the research process controlled by the investigator. Publications are considered a sufficient public record of the research, subject to scrutiny by peers. Data could be discarded within some reasonable period of time after findings were published. Retaining proprietary control over data is powerful; those data can be reused by investigators and can be bartered for other data, for collaborators, and for funding. The open access assumptions in the policy maker's scenario treat data more as public goods; they are assets to be released in return for public funding. Transparency is also assumed to reduce fraud and unethical behavior.

Somewhere in the middle of these competing scenarios and assumptions is the need to govern data – whatever they are – and their uses. Rarely are research data true public or true private goods. More often they are "common-pool resources" whose control is sufficiently contentious that they need to be governed (Borgman, 2015; Hess & Ostrom, 2007).

Each of the three scenarios contains truths with which the community must reckon. Data are assets to be managed in the short and long term. Many kinds of data have value for other purposes. Documenting data adequately to make them reusable for others is a complex and expensive task that requires considerable expertise. Few researchers have the necessary skills or resources to invest in stewardship of their data. Few universities are investing substantial resources in local data archiving. Social science data archives

such as ICPSR serve essential roles in the knowledge infrastructure. They must be nurtured and sustained, but substantial expansion of such institutions will be required if social science data are to be sustained at scale. The next generation of researchers must learn not only modern research methods, they must learn modern methods of data management if they are to exploit their data assets effectively over the course of their careers. A new generation of data scientists who can work with researchers, data archives, libraries, and other stakeholders to steward data assets also is necessary.

Research data have much in common with the song, *Not Fade Away,* with lyrics by Charles Hardin and Norman Petty, presented above as an epigraph. The origins are hard to trace and the variants are many, each with new interpretations. Those who know *Not Fade Away* as a Grateful Dead song will resent those who attribute it to the Rolling Stones. Buddy Holly fans will take umbrage at both. Only music librarians and diligent Wikipedia searchers will know that Buddy Holly's birth name was Charles Hardin Holley and that he sometimes published under his first two given names. Unless such disputes can be set aside in favor of a serious discussion about how to govern research data in the social sciences, those data will fade away. Your love for data has got to be real.

References

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: The MIT Press.

Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Recommendation for space data system practices No. CCSDS 650.0-M-2 Magenta Book). Washington, D.C.

Hardin, C., & Petty, N. (1957). Not Fade Away Lyrics. Retrieved April 15, 2016, from http://www.metrolyrics.com/not-fade-away-lyrics-grateful-dead.html

Hess, C., & Ostrom, E. (2007). *Understanding knowledge as a commons: from theory to practice*. Cambridge, MA: MIT Press.