Harvard Data Science Review • 1.1

The Lives and After Lives of Data

Christine L. Borgman

Published on: Jun 23, 2019

Updated on: Oct 04, 2019

DOI: 10.1162/99608f92.9a36bdb6

ABSTRACT

The most elusive term in data science is 'data.' While often treated as objects to be computed upon, data is a theory-laden concept with a long history. Data exist within knowledge infrastructures that govern how they are created, managed, and interpreted. By comparing models of data life cycles, implicit assumptions about data become apparent. In linear models, data pass through stages from beginning to end of life, which suggest that data can be recreated as needed. Cyclical models, in which data flow in a virtuous circle of uses and reuses, are better suited for irreplaceable observational data that may retain value indefinitely. In astronomy, for example, observations from one generation of telescopes may become calibration and modeling data for the next generation, whether digital sky surveys or glass plates. The value and reusability of data can be enhanced through investments in knowledge infrastructures, especially digital curation and preservation. Determining what data to keep, why, how, and for how long, is the challenge of our day.

Keywords

astronomy, curation, data, digital curation, life cycles, observations, preservation, reuse, science, stewardship

1. Introduction

As an interdisciplinary journal of data science whose goal is to provoke dialog among diverse stakeholders, the *Harvard Data Science Review* is an ideal venue to explicate concepts whose terminological simplicity masks highly contested territory. 'Data' is the most elusive term of all. Data are often treated as objective entities to be computed upon, defined as facts or numbers, or operationalized by lists of examples. In practical business situations where correlation matters more than causation, such declarative simplicity may suffice. In scholarly contexts, however, data, facts, information, and knowledge are theory-laden concepts with long and contentious histories (Blair, 2010; Buckland, 1991; Case, 2006; Leonelli, 2015; Meadows, 2001; Rosenberg, 2013). Researchers are exceedingly clever at treating almost anything as data, be it the air we breathe, clothes we wear, traces of our digital lives, or photons captured by astronomical instruments. In scientific contexts, data can be viewed as "entities used as evidence of phenomena for the purposes of research or scholarship" (Borgman, 2015, p. 29). From a humanities perspective, "the concept of *data* as a given has to be rethought through a humanistic lens and characterized as *capta*, taken and constructed. … rooted in a co-dependent relation between observer and experience" (Drucker, 2011).

2. Data and Infrastructure

Whether in science, humanities, business, or government contexts, data are a human construct. People decide what are data for a given purpose, how those data are to be interpreted, and what constitutes appropriate evidence. One scientist's signal is another's noise. One politician's fact is another's fake news. Data exist within knowledge infrastructures that govern how they are created, managed, used, and interpreted (Edwards et al., 2013). As infrastructures evolve, so do the characteristics and usability of data embedded within them.

The notion of 'data life cycle' reflects the array of knowledge infrastructures that govern the flows of data. The term *life cycle* originated in biology in the 19th century as a linear model ("Oxford English Dictionary," 2019): "The sequence of stages through which an individual organism passes from origin as a zygote to death, or through which the members of a species pass from the production of gametes by one generation to that by the next." Life cycle is used similarly in business and economic contexts to span processes from their beginning through decay or ending. An example is personnel records that are created when a person is hired and destroyed at the end of a legally defined records retention cycle.

The common alternative to a linear data life cycle is a circular model, where data flow continually through stages. These models are common in scholarly communication and in other areas that benefit from the ability to mine and combine data indefinitely. Figure 1, a 'research life cycle' from a library perspective, illustrates the flow of scholarly products. In the planning stage of a project, researchers typically describe a problem and determine the research design. In the implementation stage, assets such as data are collected, organized, described, and analyzed. The next stage is to publish the resulting work, which may include depositing associated datasets for public access. Once published, the research findings may be disseminated further through social media, indexing and abstracting services, and various 'impact' mechanisms. The next stage in Figure 1 is preservation, which includes reliable storage and migration to new technologies that ensure continuous availability. The last and connecting stage is reuse, when research products become input to the planning and implementation of new research projects.

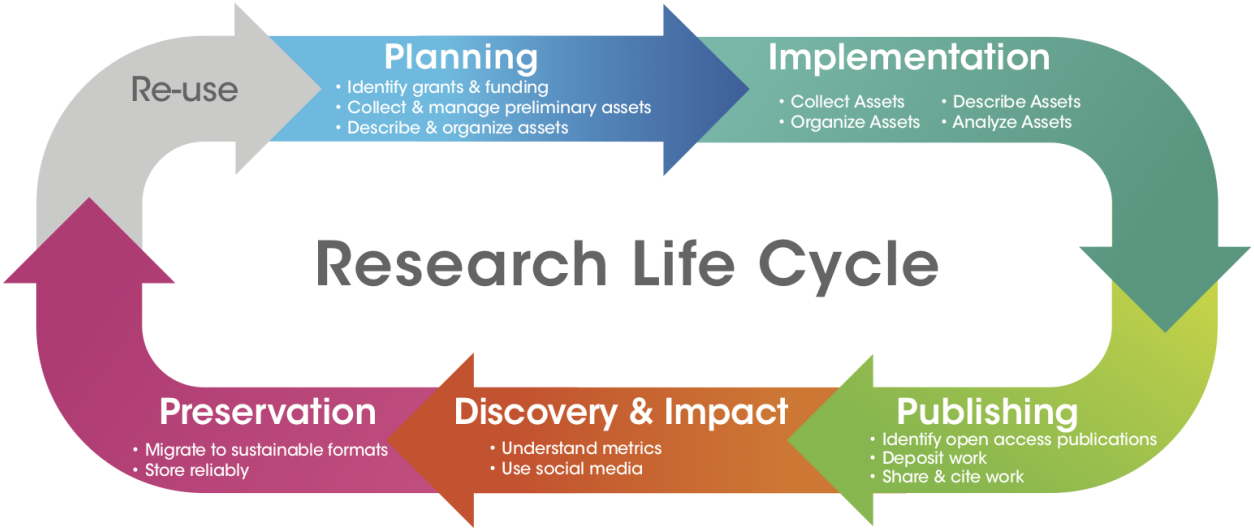

Figure 1: Research Life Cycle (University of California, Irvine, Libraries, Digital Scholarship Services, 2019). Reprinted with permission of the UCI Libraries.

The idea behind the life cycle model in Figure 1 is to encourage researchers to think in terms of a virtuous circle wherein their work has greater impact, for longer periods of time, through dissemination and preservation of their research products. Libraries provide essential elements of the knowledge infrastructure for this virtuous circle, such as dissemination, curation, preservation, and access. In principle, a student or other researcher could begin an inquiry at any point in the cycle or could skip a stage or two. Questions provoked by the dissemination process could lead to reuse of data, as could datasets stored in archives, for example. Conversely, projects may proceed only through parts of this research life cycle. Researchers may fail to complete a project or fail to publish their findings. Publications may or may not receive citations from other authors. Only a minority of researchers preserve their datasets in ways that the data remain findable and accessible. Even if datasets are available, those data may not be reused by others.

Figure 2—a much more complex model that is widely adopted in the digital archiving community—also focuses on keeping digital data alive for long periods of time. Books and other paper objects often can survive indefinitely by benign neglect, given adequate storage conditions. Digital records, in contrast, require active management. The digital curation life cycle model in Figure 2, explained more fully in Higgins (2008) and on the DCC site (Digital Curation Centre, 2019), identifies activities that keep data available, useful, and usable. During reappraisal, archivists determine whether to continue investment in a dataset, such as migrating it to new formats and media, or whether to dispose of the dataset. Digital data archives of scholarly content, such as ICPSR in the social sciences, GBIF for

biodiversity, UniProt for protein sequences, HEASARC for high energy astrophysics, or DANS for humanities and archaeology, all invest in data curation in a manner similar to that of the DCC model (Data Archiving and Networked Services, 2017; GBIF, 2019; "HEASARC: NASA's Archive of Data on Energetic Phenomena," 2019; "Inter-university Consortium for Political and Social Research," 2019; "UniProt," 2019). Lacking these investments in data curation and preservation, data fade away through neglect, benign or otherwise, as storage media fail and as software versions become obsolete (Borgman, 2015, 2016).

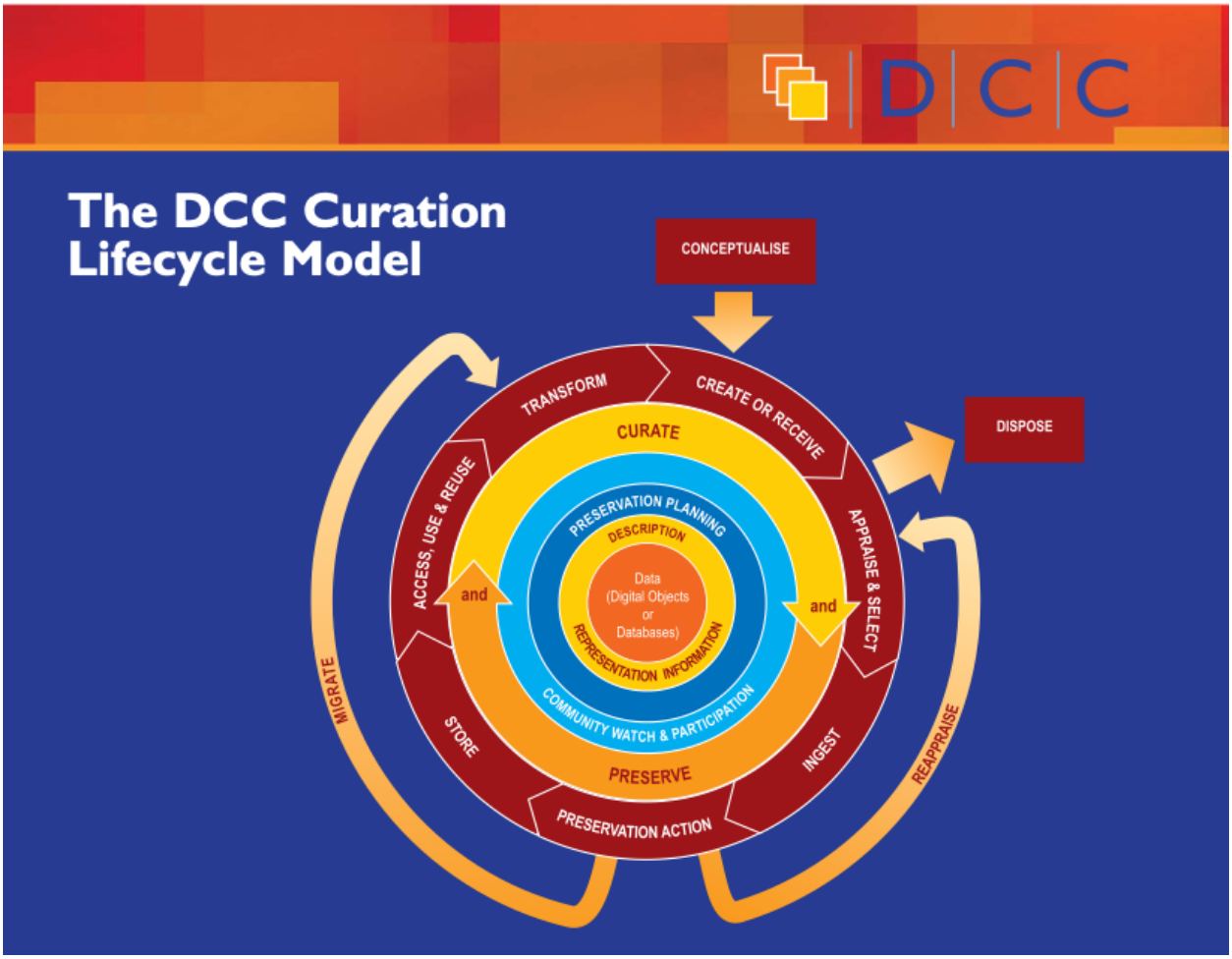

Figure 2: Digital Curation Center Curation Lifecycle Model (Higgins, 2008). Reprinted with permission of the Digital Curation Centre, U.K.

The stark contrast between the popularity of linear life cycles in technical areas of data science and cyclical life cycles in the digital curation community reveals competing assumptions about data and infrastructure. If data exist only from the time they are generated *de novo* to when they are interpreted (Wing, 2018; Wing, Janeja, Kloefkorn, & Erickson, 2018), they are ephemeral objects produced for a specific purpose. They can be discarded without further investment. In contrast, if data are entities humans created as evidence of a particular phenomenon, they may have enduring value. If

those data are to be reused, they must be reusable, which requires considerable investment in the infrastructure necessary for documentation, interpretation, curation, and access.

Another implicit assumption about data that distinguishes these life cycle models is whether data can be recreated. Experiments and computational models can be re-executed, social media streams can be resampled, and even genome sequences can be recreated if the original tissue is available and viable. Observational data, in contrast, cannot be recreated. The census of 2010 cannot be conducted again, nor can infrared images of tonight's sky be taken tomorrow, nor can the weather conditions of July 4, 1776, be observed again with modern instruments. These are time-specific observations that may be valuable indefinitely. One never steps in the same river twice, because the water continues to flow. That said, not all observational data can be kept alive, nor are all worth keeping.

3. Open Science and Data Stewardship

Research policy initiatives for open science, open access to publications, data management plan requirements, and deposit of data associated with publications are predicated on assumptions that research data are valuable assets that should be preserved for reuse by others, whether for reproducibility, reuse for new questions or innovations, mining and integration, or other purposes (National Academies of Sciences, 2018; "NumFOCUS," 2018; Wilkinson et al., 2016). Implicit in these policies are assumptions that research data should be curated and preserved to become part of the virtuous circle presented in Figure 1.

Astronomy offers numerous examples of cyclical data life cycles in which reuse is essential, as each round of observations and instrumentation lays the foundation for the next. Human observations of the cosmos long predate the written record, and the cosmos long predates humans. A contemporary case to consider is the Large Synoptic Survey Telescope (LSST), which is in its final stages of construction in Chile. "Engineering first light" is due in FY 2020 and science operations are due to begin in FY 2021, commencing 10 years of data collection (Ivezic et al., 2008; Large Synoptic Survey Telescope, 2019). Many milestones could be chosen to mark the beginning of LSST. Concept development and proposals began in the 1990s, long before funding for the telescope instrument was obtained. Countless design decisions and compromises were made by the time the glass was poured for the mirror, thus hardening the path to data collection. Many of these design decisions are based on data obtained by earlier surveys and instruments. Observations from the Sloan Digital Sky Survey, a ground-based survey that saw first light in 1998 and entered routine operations in 2000 ("Sloan Digital Sky Surveys," 2019), are among those used to calibrate LSST.

More than half of the one billion dollar budget of the LSST project is devoted to data management because those data are expected to remain valuable to several generations of astronomers. The science

is in the data. Major astronomy missions such as Chandra and Hubble report that more new papers are being published from their archival data than from new observations ("Chandra Data Archive," 2019; "Hubble Legacy Archive," 2019).

Old observational data yield new forms of evidence and new baselines for current evidence. LSST is expected to benefit greatly from DASCH, a project begun in 2005 to digitize the Harvard Observatory's collection of a half-million glass plates, acquired over a period of more than a century. Because the irreplaceable observations captured on these plates represent the first complete map of the sky, they are an essential baseline comparison for LSST and other sky surveys. The scientific value of DASCH lies in the infrastructure that encompasses carefully curated data, high resolution imaging, and computational features that enable astronomers to explore and visualize time-domain astronomy in ways inconceivable when these data were collected in the 19th and 20th centuries (Digital Access to a Sky Century @ Harvard, 2019; Grindlay, Tang, Los, & Servillat, 2011; Sobel, 2017).

The lives and afterlives of data depend upon many factors, such as their perceived value and the efforts invested in their curation. Glass plates fell into disuse for scientific purposes when charge-coupled devices (CCDs) became a viable technology. These plates are large and fragile objects that are expensive to maintain, and thus many were discarded by the time that astronomy became digital. Harvard, despite the continuing specter of fires, floods, and budget cuts, managed to keep their plate collection and catalogs intact. The dedication of a core group of individuals facilitated the digital archive that is now openly available to the international community.

4. Knowledge Infrastructures for the Long Term

Data life cycles, whether viewed as linear or cyclical processes, are necessarily reductionist. Paths from data creation to interpretation and back tend to look more like a random walk than a perfect line or circle. Infrastructures, by their nature, tend to be most visible when they break down. They build on an installed base and are embedded in the social practices of their communities (Star & Ruhleder, 1996). Data are selected, collected, organized, and generated by humans, using the knowledge infrastructures available to them at the time. Some of those data may be short-lived, discarded when they have served their purpose, and readily recreated if later needed. Other data, such as observations of the natural world, may be long-lived, with value apparent from their initial capture. Much else falls in between, including observations lost before their value was recognized, duplicative material that can be done without, and sensitive data that should be destroyed regularly due to privacy and ethics risks. In data science, we ignore knowledge infrastructures at our peril. Identifying principles for what to keep, why, how, and for how long, is the challenge of our day.

Acknowledgements

Research on astronomy data practices reported here was supported by the Alfred P. Sloan Foundation, *If Data Sharing is the Answer, What is the Question*?, Sloan #2015-14001, Christine L. Borgman, PI, and by the Harvard-Smithsonian Center for Astrophysics as a Visiting Scholar. Thanks to John P. Renaud of University of California, Irvine, Libraries for permission to use Figure 1 and to Kevin Ashley of the U.K. Digital Curation Centre for permission to use Figure 2. Bernadette Boscoe, Michael Scroggins, Morgan Wofford, and Peter Darch of UCLA Center for Knowledge Infrastructures provided comments on an earlier draft.

References

Blair, A. M. (2010). *Too Much to Know: Managing Scholarly Information Before the Modern Age*. New Haven, CT: Yale University Press.

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

Borgman, C. L. (2016, June 23). Not Fade Away: Social Science Research in the Digital Era [The Social Science Research Council]. Retrieved from Parameters website: http://parameters.ssrc.org/2016/06/not-fade-away-social-science-research-in-the-digital-era/

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science*, *42*(5), 351–360. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<351::AID-ASI5>3.0.CO;2-3

Case, D. O. (2006). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (2nd ed.). San Diego: Academic Press.

Chandra Data Archive. (2019). Retrieved January 21, 2018, from http://cxc.harvard.edu/cda/

Data Archiving and Networked Services. (2017). DANS: Organisation and policy. Retrieved July 11, 2017, from https://dans.knaw.nl/en/about/organisation-and-policy

Digital Access to a Sky Century @ Harvard. (2019). DASCH Data Release. Retrieved April 5, 2019, from http://dasch.rc.fas.harvard.edu/index.php

Digital Curation Centre. (2019). Retrieved November 15, 2016, from http://www.dcc.ac.uk/

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, *005*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., … Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges* (p. 40). Retrieved from University of Michigan website: http://hdl.handle.net/2027.42/97552

Global Biodiversity Information Facility. (2019). Retrieved April 24, 2019, from https://www.gbif.org/

Grindlay, J., Tang, S., Los, E., & Servillat, M. (2011). Opening the 100-Year Window for Time-Domain Astronomy. *Proceedings of the International Astronomical Union*, *7*(S285), 29–34. https://doi.org/10.1017/S1743921312000166

HEASARC: NASA's Archive of Data on Energetic Phenomena. (2019). Retrieved April 24, 2019, from https://heasarc.gsfc.nasa.gov/

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Hubble Legacy Archive. (2019). Retrieved April 25, 2019, from http://hla.stsci.edu/

Inter-university Consortium for Political and Social Research. (2019). Retrieved October 6, 2016, from https://deepblue.lib.umich.edu/handle/2027.42/57738

Ivezic, Z., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., AlSayyad, Y., … Collaboration, for the L. (2008). LSST: from Science Drivers to Reference Design and Anticipated Data Products. *ArXiv:0805.2366 [Astro-Ph]*. Retrieved from http://arxiv.org/abs/0805.2366

Large Synoptic Survey Telescope. (2019, March 28). LSST Project Schedule. Retrieved March 28, 2019, from https://www.lsst.org/about/timeline

Leonelli, S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, *82*(5), 810–821. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4747116/

Life cycle. (2019). In *Oxford English Dictionary* (Online). Retrieved from www.oed.com

Meadows, J. (2001). *Understanding Information*. Munchen: K. G. Saur.

National Academies of Sciences, E. (2018). *Open Science by Design: Realizing a Vision for 21st Century Research*. https://doi.org/10.17226/25116

NumFOCUS: Open Code = Better Science. (2018). Retrieved October 22, 2018, from NumFOCUS website: https://numfocus.org/

Rosenberg, D. (2013). Data before the Fact. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 15–40). Cambridge MA: MIT Press.

Sloan Digital Sky Surveys. (2019). Retrieved March 28, 2019, from https://www.sdss.org/surveys/

Sobel, D. (2017). *The Glass Universe: How the Ladies of the Harvard Observatory Took the Measure of the Stars* (Reprint edition). New York: Penguin Books.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

UniProt. (2019). Retrieved April 24, 2019, from https://www.uniprot.org/

University of California, Irvine, Libraries, Digital Scholarship Services. (2019). Research Life Cycle. Retrieved February 15, 2019, from https://www.lib.uci.edu/dss

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. Retrieved from http://dx.doi.org/10.1038/sdata.2016.18

Wing, J. M. (2018, January 23). The Data Life Cycle | Data Science Institute. Retrieved March 28, 2019, from https://datascience.columbia.edu/data-life-cycle

Wing, J. M., Janeja, V. P., Kloefkorn, T., & Erickson, L. C. (2018). *Data Science Leadership Summit: Summary Report*. USA: National Science Foundation.
