Context and Collection:
A Research Agenda for Small Data

Amelia Abreu
University of Washington
ameliacabreu@gmail.com

Amelia Acker
University of California, Los Angeles
aacker@ucla.edu

Abstract

In recent years, *big data* has become a prevalent issue for information studies. In an era of big data, can we contemplate research data that relies more on the context of creation than volume and variety of source. In this note, we report on early findings of phenomena we identify as *small data.* Despite the outpouring of critique and theoretical assertions related to big data, little attention has been paid to the collections, researchers and collecting institutions that get left out the rhetoric of big data. We present criteria for small data and explore some of the issues inherent in developing small data research. The resulting analysis develops future directions towards a comprehensive small data research agenda. We also develop and discuss factors for consideration in context, preservation and access of both big and small data.

Keywords: Big data, data curation, infrastructure studies, small data

Introduction: What Big Data Leaves Behind

In recent years, claims of a big data revolution have generated significant response in information studies. Big data has been positioned as both a must-have commodity and resource in business, government, academic research, and military applications [1–3]. Oracle, Microsoft, and Intel have each developed big data analytic tools for enterprise [4–6]. Likewise, academic big data initiatives in disciplines such physics and astronomy have promised new levels of analytics and discovery. Even the National Endowment for the Humanities has created the "Digging into Data Challenge," partnering with international collaborators to challenge social scientists and humanists to incorporate large data corpora and data intensive computational techniques into their research agendas [7], [8]. We have the opportunity to consider context, value and research methods now that the turn towards big data has been identified.

In this paper, we define the phenomena of *small data* as it relates to big data narratives in information science and social science research. Critics have identified how big data will create new digital divides in analysis of and access to data, as well as tools and levels of expertise that can be applied once it has been gathered [9], [10]. How do new contexts of big data subsume the possibilities for rich and faceted interpretation of data that is 'too small' to fit in the current discourse of processing power, distributed computing resources, walled platforms, and data analytics?

Small data*,* as we refer to it, exists as a growing area of study that has been overlooked in the big data ecosystem. By engaging with small data, we may critique the assumptions, processes and commitments of big data. Because big data is seen as an extension of the data-intensive, fourth paradigm of science, it provides information scholars the opportunity to look closely at the places where small data lives, is used for discovery, and has been preserved for access in decades past.

In this note, we propose the following questions for small data research in the era of big data:

- What is "small data" and how does it differ from big data?
- How can the value of small data research be articulated in response to the drive for big data?
- How can we design tools and information systems for small data?

Abreu, A., & Acker, A. (2013). Context and collection: A research agenda for small data. *iConference 2013 Proceedings* (pp. 549-554). doi:10.9776/13275

Related Work

In developing a small data research agenda, we draw on two main areas of work: critiques of big data research and possible tactics for small data approaches, located in the study of infrastructure, the digital curation movement and personal digital archiving.

Critiques of Big Data

While business, scientific and social research are using big data to drive change and ask new questions, some critics have expressed shortcomings. The most common critique of big data has been the question of access to data and in turn the possibilities for analysis. Manovich has coined these access issues as a "data analysis divide" [10]. One main point of contention is that of agency in relation to personal information. As users contribute content and data in social media platforms, it remains to be seen who gets access to these data sets and under what auspices. The majority of big data research takes place in the environments with the most resources, either in corporate settings or top-tier research universities. Furthermore students and faculty from such universities are more likely to gain access to corporate research settings and granting agencies, and thus "set[ting] up new hierarchies about 'who can read the numbers'" [9].

Another problem with big data applications is the range of the possible social scientific questions that can be asked of it in combination with scale, or what some communication scholars have called 'internet time' [12–14]. Big data sources collected online are often incomplete because of closed-platforms, privacy policies, speed of change, and limited information access. Many of the research methods that are being applied to big data are untested. David Karpf has written about some of these methodological cul-de-sacs that are results of "endemic problems associated with online data quality," noting that researchers who work with big data are "well aware of its limitations. Spambots, commercial incentives, proprietary firewalls, and noisy indicators all create serious challenges." [12].

Big data are not isolated data sets, cordoned information, subject or platform specific data. As danah boyd and Kate Crawford argue, the most important aspect of big data is the possibilities for relationality with other data: "the value is in making connections between sets" [16]. Corporations have developed big data management services for scraping and cleaning unstandardized data for storage, access, and analysis [4–6]. Relating data sets is only possible if the owners, creators, and generators/producers of data have the means to do so.

Infrastructure studies, Data Curation, and Personal Digital Archiving

The convergence of big data practice and intensive-data discovery in scholarship and professional practice in information science can be seen in the areas of infrastructure studies, the recent history of data curation and the emerging field of personal digital archiving (PDA). Geoffrey Bowker and S. Leigh Star have written extensively about the value-laden and performative aspects of infrastructuring that are involved when heterogeneous data-sets are combined in data-intense science [17–19]. By analyzing how values are ascribed in the structure and layering of technology, protocols and standards, we may locate the performative nature of databases and our abilities to access that data in different ways. We can locate how big data enables new ways of knowing by employing tactics from infrastructure studies [20]. Additionally, by being attuned to change across information infrastructures we may also locate *how we have known* in the past, and further the ways of knowing with small data before and during the big data era.

Data curation techniques focus upon the quality, trustworthiness, re-use of data for discovery, and they are often framed as confronting the 'data deluge' [21–23]. Small data research agendas can build upon data curation techniques that rely upon the continuous enrichment of data; including, starting from the contexts of creation, a commitment to metadata and a deep understanding of its shifting, often ephemeral qualities [24], [25].

Another area ripe for small data scholarship can be found in studies that take seriously the personal digital archiving practices of individuals documenting their digital lives [26–28]. While description

and expectations for future use are recurring themes in narratives of big data, they also have long history in archival studies and practitioner research focused on enduring value and stewardship [29], [30]. Engagement with personal digital archiving is another means of getting at the nuances and variety of possible small data research questions and applications.

While the "big data rich" can be found in industry, government, and research universities, the small data rich may be all around us, in the cultural heritage institutions like libraries, museums and archives that are committed to public access and accountability [9], [31]. Moreover, humanistic, historical and social science methods emphasize bounding research data that privileges context, the process of creation and capture, and emphasizes principles of preservation.

Researching small data begins by countering big data's hype.. The study of infrastructure, the expertise and analytics of data curation, as well as the turn towards personal records management in PDA each confront the nuances, half-life, affects and influences of the big data era.

Defining Small Data

In order to define small data, we must first articulate its distinctions. In this section, we posit six central distinctions of small data: *motivation*, *data collection*, *context*, *affect*, *archival engagement*, and *retention*. By doing so, we develop an operational set of factors for analysis and discussion (Figure 1.).

Small Data	Big Data
• Data collected purposefully in defined settings • Data collection is local: conducted by individuals or teams • Data collection is held to disciplinary and community standards • Data set is part of research archive and made visible to community • Institutional, professional, and disciplinary standards for rigor and ethics in data collections • Human labor in data collection as professional or scholarly activity • Informed consent of subjects	• Data collected transactionally in a variety of settings, across platforms • Data collection is automated: conducted by retrieval tools • Data collection not standardized: dependent on setting, platform, and terms of service • Data set is proprietary, access limits • Lack of standards for rigor, quality and ethics in data collection • Data collection is de-professionalized labor • Implied consent of subjects

Figure 1. Key Distinctions between Big and Small Data.

To begin, small data is collected differently than big data, usually with defined parameters and boundaries. Instead of collecting data indiscriminately and automatically, small data collecting motivations are articulated at the outset. These motivations are generally purposeful and to a certain degree, conscientious, stemming from a research question, a hypothesis, or an individual or group mandate.

Moreover, the professional, scholarly, or cultural mandate to collect data holds the work to a higher standard than that of big data collection, with considerations for subjects, context, and impact of data collection reflecting an established ethic and rigor. In Jenna Burrell's example of ethnographic data, the research data set documents "some kinds of things straightforwardly, but not others" [15]. In scholarly communities such as ethnography, data collection is professional work of utmost importance.

Context is perhaps the most important factor that distinguishes the small data approach, as it is the most difficult factor to regain when lost. For ethnomusicologists, fieldwork and data collection "should

happen where music happens" [32], and thus an ethnomusicologist's data set consists not only of music, but other observations as to its context and "happening". In this community, big data methods "can at once benefit our work and magnify our reflexive anxieties about the impact of our data collection on our ethnographic integrity" [32].

Small data collection requires a deeper *affective* engagement, both in terms of professional labor and personal commitment. For ethnomusicologists and other scholars interested in the contextual and affective qualities of their research data (such as art historians or literary scholars), existing big data tools cannot parse, analyze, or comprehend the significance of their data sets with the level of critical rigor these types of inquiry demand. Although big data research has begun to venture into affect inquiry [33], the resulting research is dramatically different than that of traditional scholarship in the field [34].

In the humanistic tradition, use of data as texts, artifacts, or observations is seen as an act of *archival engagement*, one deeply ingrained in the values and ethics of scholarship. Literary scholar Lauren Berlant recounts a colleague's comment, "I hate your archive," in relation to the texts and films she presented for analysis in a talk. Berlant asks, "Was this an aggressive disciplinary question?" pointing out the affective dimensions selection, collection and interpretation of such data for humanists [35].

Small data research demands not only critical attention towards the collection process, but also for standards of *retention*. Across disciplinary lines and research settings, scholars have established different retention standards to ensure quality, respect privacy, and provide access to their data/research process. For example, historians cite publicly available records; while anthropologists are expected to anonymize field notes; while poets use manuscript drafts as part of a literary archive [36]. In big data research, retention and custody of data sets remains an unsolved and often controversial issue [37].

Small Data Futures

It is highly likely that small data will intersect with big data in the coming years, and that small data and big data research will overlap in complementary praxis. Hybrid models of integrating big data with small data have emerged, such as Anderson et al's work [38] on integrating sensor data with ethnographic fieldwork, and Batty's [39] agenda for human geography data. For many researchers seeking to integrate big data methods, a key challenge is making big data smaller: scaling available data to the parameters of the focus of research questions. As we go forward, we anticipate future research in both design for small data and policy for its access and preservation.

As we have argued, small data is remarkably dependent on context, the future of small data research must be innovative in this regard. The contributions of infrastructure studies, data curation and personal digital archives can aid in developing a more nuanced small data model. While some may argue that big data will usher in an "end of theory" [41] scholars such as Nigel Thrift [42] and Donna Haraway [43] offer valuable theoretical contributions for framing personal
and social phenomena in the big data environment. Indeed, the persistence of small data shows that life in the era of big data is complex, but that individual agency is both possible and necessary.

References

[1] C. Borgman, J. Wallis, and N. Enyedy, "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries," *International Journal on Digital Libraries*, vol. 7, no. 1, pp. 17–30, 2007.

[2] C. Lynch, "Big data: How do your data grow?," *Nature*, vol. 455, pp. 28–29, Sep. 2008.

[3] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and Kruschwitz, "Analytics: The New Path to Value," MIT Sloan Management Review, Massachusetts, 2010.

[4] "Oracle Big Data Appliance." [Online]. Available: http://www.oracle.com/us/products/database/big-data-appliance/overview/index.html [Accessed: 10-Sep-2012].

[5] "Big Data Analytics | Microsoft SQL Server." [Online]. Available: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/big-data-solution.aspx [Accessed: 10-Sep-2012].

[6] “Big Data-Intelligence Begins with Intel.” [Online]. Available: http://www.intel.com/content/www/us/en/big-data/big-data-analytics-turning-big-data-into-intelligence.html [Accessed: 10-Sep-2012].
[7] “Digging Into Data Challenge | National Endowment for the Humanities.” [Online]. Available: http://www.neh.gov/grants/odh/digging-data-challenge [Accessed: 10-Sep-2012].
[8] C. Williford, C. Henry, and A. Friedlander, “One Culture: Computationally Intensive Research in the Humanities and Social Sciences,” Council on Library and Information Resources, Washington, D.C., pub 151, 121AD.
[9] danah boyd and K. Crawford, “CRITICAL QUESTIONS FOR BIG DATA,” *Information,Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012.
[10] L. Manovich, “Trending: The Promises and the Challenges of Big Social Data,” in *Debates in the Digital Humanities*, M. K. Gold, Ed. Minneapolis, MN: U of Minnesota Press, 2012.
[11] D. M. Zorich, G. Waibel, R. Erway, D. M. Zorich, G. Waibel, R. Erway, O. Programs, and G. Waibel, *Beyond the Silos of the LAMs: Collaboration Among Libraries, Archives and Museums*. 2008.
[12] D. Karpf, “SOCIAL SCIENCE RESEARCH METHODS IN INTERNET TIME,” *Information, Communication & Society*, vol. 15, no. 5, pp. 639–661, 2012.
[13] B. D. Loader and W. H. Dutton, “A DECADE IN INTERNET TIME,” *Information, Communication & Society*, vol. 15, no. 5, pp. 609–615, 2012.
[14] L. A. Lievrouw, “THE NEXT DECADE IN INTERNET TIME,” *Information, Communication & Society*, vol. 15, no. 5, pp. 616–638, 2012.
[15] J. Burrell, “The Ethnographer’s Complete Guide to Big Data: Conclusions (part 3 of 3),” *Ethnography Matters*. [Online]. Available: http://ethnographymatters.net/2012/06/28/the-ethnographers-complete-guide-to-big-data-part-iii-conclusions/ [Accessed: 10-Sep-2012].
[16] danah boyd and K. Crawford, “Six Provocations for Big Data,” *SSRN eLibrary*, Sep. 2011.
[17] G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences*, 1st ed. The MIT Press, 1999.
[18] G. C. Bowker, “Biodiversity Datadiversity,” *Social Studies of Science*, vol. 30, no. 5, pp. 643–683, Oct. 2000.
[19] Star, Susan Leigh and Bowker, Geoffrey C., “How to Infrastructure,” in *Handbook of new media : social shaping and social consquences of ICTs*, London, Thousand Oaks: SAGE Publications.
[20] Bowker, Geoffrey C., K. Baker, F. Millerand, and D. Ribes, “Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment,” in *International Handbook of Internet Research*,
J. Hunsinger, L. Klastrup, and M. Allen, Eds. New York: Springer, 2010.
[21] H. Karasti, K. S. Baker, and E. Halkola, “Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network,” *Computer Supported Cooperative Work (CSCW)*, vol. 15, no. 4, pp. 321–358, Aug. 2006.
[22] G. Bell, T. Hey, and A. Szalay, “COMPUTER SCIENCE: Beyond the Data Deluge,” *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 2009.
[23] P. Lord, A. Macdonald, L. Lyon, and D. Giaretta, “From Data Deluge to Data Curation,” in *In Proc 3th UK e-Science All Hands Meeting*, 2004, pp. 371–375.
[24] J. Gray, A. S. Szalay, A. R. Thakar, C. Stoughton, and J. vandenBerg, “Online Scientific Data Curation, Publication, and Archiving,” *arXiv:cs/0208012*, Aug. 2002.
[25] N. Beagrie, “Digital Curation for Science, Digital Libraries, and Individuals,” *International Journal of Digital Curation*, vol. 1, no. 1, Dec. 2008.
[26] N. Van House and E. F. Churchill, “Technologies of memory: Key issues and critical perspectives,” *Memory Studies*, vol. 1, no. 3, p. 295, 2008.
[27] N. Van House, M. Davis, M. Ames, M. Finn, and V. Viswanathan, “The uses of personal networked digital imaging: an empirical study of cameraphone photos and sharing,” in *CHI ’05: CHI ’05 extended abstracts on Human factors in computing systems*, Portland, OR, USA, 2005, pp. 1853–1856.
[28] C. C. Marshall, S. Bly, and F. Brun-Cottan, “The Long Term Fate of Our Digital Belongings: Toward a Service Model for Personal Archives,” *arXiv:0704.3653*, Apr. 2007.

[29] C. A. Lee, *I, Digital: Personal Collections in the Digital Era*. Amer Library Assn, 2011.
[30] A. Cushing, “Highlighting the archives perspective in the personal digital archiving discussion,” *Library Hi Tech*, vol. 28, no. 2, pp. 301–312, Jun. 2010.
[31] J. Anderson and L. Rainie, “The Future of Big Data,” Pew Internet & American Life Project, Washington, D.C., 2012.
[32] T. Cooley, K. Meizel, and N. Syed, “Virtual Fieldwork: tree cases studies,” in *Shadows in the Field: New Perspectives for Fieldwork in Ethnomusicology*, Oxford University Press, USA, 2008, pp. 90–107.
[33] A. D. I. Kramer, “The spread of emotion via facebook,” in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 767–770.
[34] L. Grossberg, *Cultural Studies in the Future Tense*. Duke University Press, 2010.
[35] L. Berlant, *The Queen of America Goes to Washington City: Essays on Sex and Citizenship*. Duke University Press, 1997.
[36] S. J. Jackson, P. N. Edwards, Bowker, Geoffrey C., and C. P. Knobel, “Understanding infrastructure: History, heuristics and cyberinfrastructure policy,” *First Monday*, vol. 12, no. 6, Jun. 2007.
[37] D. Bollier, *The Promise and Peril of Big Data*. Aspen Institute, Communications and Society Program, 2010.