THEODORE CALVIN PEASE AWARD

Big Questions: Digital Preservation of Big Data in Government

Emily Larson

ABSTRACT

Big Data is becoming a key part of transactions and decision-making processes, and archivists are increasingly called to intervene in its management. This article examines the digital preservation needs of government Big Data from the perspective of archival theory. While Big Data presents unique challenges, particularly in the areas of record capture, access, and privacy, it is nonetheless becoming a key component of modern government recordkeeping. Managing both the technical and ethical aspects of Big Data is essential, with each requiring specific consideration. Taking a systems-level view of Big Data by attempting to capture instances of bounded variability may be one path forward, and technical tools and systems can successfully manage such large volumes of information. However, ultimately, as with all digital preservation initiatives, proper documentation is key. Creating appropriate metadata to capture the identity, technical characteristics, and management actions for Big Data must include the multiprovenancial origins of such data sets. More broadly, Big Data reminds archivists of their larger responsibilities. Recognizing the power dynamics in Big Data requires an interrogation and documentation of the data themselves, as well as of the ways in which governments and corporations use them. Digital preservation must balance technical knowledge with critical perspectives to truly capture the context of Big Data and the records it produces.

© Emily Larson.

KEY WORDS

Digital preservation, Big Data, Government archives, Privacy, Ethics, Digital archives, Data repositories

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Big Data is becoming a key part of transactions and decision-making processes, and archivists are increasingly called to intervene in its management. Should they? When governments adopt Big Data strategies, the resulting transactional records have inherent bonds to the data they are based upon. Big Data is sometimes defined as an archives in and of itself, but these claims often emphasize storage over archival principles, particularly authenticity and long-term preservation. This article examines the digital preservation needs of government Big Data from the perspective of archival theory. While Big Data presents unique challenges, particularly in the areas of record capture, access, and privacy, it is nonetheless becoming a key component of modern government recordkeeping. Managing both the technical and ethical aspects of Big Data is essential, with each requiring specific consideration. Archivists are well positioned to intervene in the creation and use of Big Data to support citizens' rights and government accountability.

Discussions of Big Data have skyrocketed in recent years, but what exactly is Big Data? Initially, Big Data referred to "data sets large enough to require supercomputers."[1] However, as computational power became more accessible, Big Data was defined in terms of "the 3Vs: volume, velocity and variety."[2] Volume refers to the large quantities of data, velocity indicates the rapid proliferation and "temporal dynamism" of data, and variety reflects the heterogeneous nature of Big Data.[3] Rob Kitchin expands upon the 3Vs to add that Big Data is "exhaustive in scope"; "fine-grained in resolution"; "relational in nature"; and "flexible."[4] Dynamism, diversity, and analytic power are essential characteristics of Big Data. Now, Big Data commonly includes a fourth V: "veracity," which speaks to the "uncertainty of data."[5] Mistrust around the quality of data harkens back to archival concepts of accuracy, reliability, and authenticity. It also reminds us that powerful systems have real-world impacts that need to be thoughtfully considered. Analytic power in particular is central to understanding the value and usage of Big Data; danah boyd and Kate Crawford note that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets."[6] Big Data can make large, unwieldy sets of information more workable and understandable for human users. However, it is important to view Big Data insights with a critical perspective. Boyd and Crawford define Big Data in terms of technology, which reflects computational capabilities; analysis, which refers to the use of Big Data to identify patterns and make larger claims; and mythology, which indicates a critical perspective on techno-solutionism that identifies Big Data as a "higher form of intelligence . . . with the aura of truth, objectivity, and accuracy."[7] The mythology of Big Data is critical for archivists to understand, as it has clear ties with archival practice. When parsing definitions of Big Data, it is important to recognize not only its technical aspects, primarily its velocity, volume, and

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

variety, but also its veracity, which speaks to claims of knowledge production and decision-making.

The analytic power provided by the use of Big Data has resulted in its widespread adoption across sectors and disciplines. Scientific research involving large data sets has pioneered work on managing Big Data.[8] Similarly, corporations driven by the profit motive are finding new ways to generate and leverage consumer data for their benefit.[9] Governments have also turned to Big Data to inform policy decisions. Because the "state is a prime generator and user of data," it makes sense that governments would seek to leverage their preexisting assets.[10] Looking at this context is useful for parsing out archival concerns because of the specific recordkeeping responsibilities of governments. For democratic governments to be accountable and transparent, records that reflect key transactions and decisions must be properly managed and preserved. Interactions between a government and its citizens often involve data collection (e.g., persons provide information to the government in exchange for services), which means that Big Data is increasingly part of this crucial relationship. As government becomes digital, Big Data emerges as a new, but important, type of government record that requires particular care and attention.

While Big Data can be used throughout governments, a few key areas are worth emphasizing. As mass generators of data, governments often use "Big Data methodologies . . . to develop new digital channels for service delivery."[11] Big Data can help improve current services, but is also powerful in its ability "to predict . . . future needs . . . and to develop the analytical capability to develop prediction models which link policy interventions to future outcomes."[12] In other words, Big Data is a key aspect of not only supporting current government services, but also of the decision-making processes involved in setting government policies. As Kate Galloway states, "Big data—rather than simply information about citizens or service users—is now assumed integral to government activity."[13] If Big Data is becoming "integral" to such activity, then it follows that Big Data will be part of the context of government records. Big Data initiatives within governments are sometimes linked to the open data movement, as well "as sharing, reuse, open access, open government, transparency, [and] accountability."[14] Beyond potential long-term preservation obligations, incorporating Big Data into government archives supports broader goals of the open government movement and can help keep governments accountable to their citizens. Again, archivists are well positioned to contribute to the management of Big Data in government because of their relationship to documenting government activities and supporting accountability.

While Big Data can support government transparency, it also has the potential for opacity and secret government activity. Big Data offers many opportunities and benefits, but no discussion of its uptake by governments is complete

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

without recognizing the role of its use in mass citizen surveillance. Through the provision of services, governments amass large swaths of data; however, some of the data are also gathered through surveillance programs. "All states are involved in surveillance for the purposes of security, safety and crime prevention and apprehension."[15] Big Data represents a significant shift in these programs because "there has been a move to replace and extend [analog systems] with digital equivalents so that they now produce big data."[16] Big Data may be regarded as a chicken and egg situation: large quantities of data require tools to make sense of them, but now the availability of these tools creates an impetus to generate and collect more data to analyze. The implications of citizen surveillance as Big Data are wide sweeping. From whistleblower Edward Snowden's revelations about the United States National Security Agency's PRISM program[17] to President Barack Obama's successful tactics during his election campaigns,[18] the use of Big Data is an essential component of modern government, including its historic victories and scandals.[19] Big Data itself is also a source of controversy, particularly regarding privacy. Many scholars have critiqued Big Data practices for failing to sufficiently protect the private information of individuals who have become data points.[20] There are significant implications for archivists seeking to preserve Big Data. On the one hand, Big Data informs some of the major events of the twenty-first century and is embedded in day-to-day government transactions; on the other hand, it poses complex, unresolved questions around access and privacy.

Before turning to the specific problems Big Data poses to archival practice, it is important to note that Big Data is often framed as an archives in and of itself. Particularly in the sciences, discussions of "the data archive" are common.[21] This archives is rarely conceptualized in terms of archival theory; the term "archives" more often reflects the storage of Big Data, rather than its long-term preservation and authenticity. That is not to say that there is no work being done on Big Data management. As one example,[22] the Sydney-AAO (Australian Astronomical Observatory) Multi-object IFS (SAMI) Survey data archives is a system that looks at longer-term preservation and access. In addition to emphasizing "open source code, ease of maintenance, and efficient storage," SAMI also considers the need for metadata and version control of data.[23] Across academic disciplines, research data management and data curation are key issues. Many universities are developing data repositories, such as the University of Porto's Information System for the Aggregated Management of Resources and Academic Records (SIGARRA), which is integrated into other university-wide systems, such as the institutional repository.[24] Discussions about data curation often include the need for digital preservation and metadata. For example, the Interdisciplinary Earth Data Alliance (IEDA) developed a thesaurus for the data it hosts because "controlled vocabulary and data consistency are

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

crucial to facilitate . . . use [of] the data."[25] Suzhen Chen and Bin Chen recognize that "collaboration between . . . information professionals and . . . researchers becomes vital to develop and enhance metadata profiles for geoscience data."[26] However, there remains a strong emphasis on the archives as a technological tool, rather than an organization with broader goals and responsibilities.[27] Though Chen and Chen highlight the importance of information professionals in their discussion of metadata, they conceptualize Big Data digital preservation in terms of "discipline-specific repositories, institutional repositories, and commercial cloud storage systems."[28] This finding aligns with the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2 project's examination of science data portals, where "many indications that elements of a 'real' archives are present," but "explicit statements of archival management are absent."[29] Beyond natural science disciplines, the social sciences and digital humanities have also taken up Big Data, and considerations of its management.[30] However, Alex H. Poole and Deborah A. Garwood reviewed the role of librarians and archivists in Big Data digital humanities projects and found that, while there is currently little representation, a clear need exists for information professionals to lend their expertise to improving Big Data management systems.[31] An archival perspective on Big Data, with its emphasis on accountability, is particularly pertinent for governments, where Big Data needs to have a clear relationship, or archival bond, to other government records so that it can provide evidence of past acts and facts. Preserving relational context is important for Big Data because it generates supporting records that are only fully understood and meaningful when connected to the more traditional records of activities that Big Data supports.

When considering the role Big Data can play in archives, it is important to ask: is Big Data a type of record? A record is "a document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference."[32] It is clear that Big Data will not completely satisfy this traditional definition. It lacks the fixity and stability of a document[33] and is often not formally set aside for the future. However, InterPARES 2's work expands the concept of a record in the digital environment "to suggest that individual digital components, or aggregations of digital components, might themselves constitute a record or a set of records, depending on how they are instantiated in the system and how they are used by the creator."[34] InterPARES 2 further adds the concept of interactive, experiential, and dynamic records.[35] Given Big Data's capacity to change based on user interaction and its dependence upon updating external or internal data sources, it aligns most closely with interactive[36] and dynamic[37] records. However, given that interactive and dynamic records require fixed rules that allow for their re-creation under the same conditions, not all Big Data is inherently an instantiation of records.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Instead, using interactive and dynamic record types as models allows the consideration of system requirements for bounding Big Data as records.

It is pertinent to think about Big Data as a type of record because of its relationship to "the course of a practical activity." Galloway argues that Big Data "has increased the scale, speed and complexity of data collection and use to such an extent that it is . . . qualitatively different from any analogue government record-keeping that has gone before it."[38] Recordkeeping in a Big Data context is different because Big Data itself informs the shape of government activities in substantially new ways. Chen and Chen argue that "Big Data is not only the collective sum of small data but can help achieve far greater contributions to research than all its parts; more importantly, it may generate new findings, applications, and solutions that cannot be possible from its constituent subsets of data."[39] Leveraging the unique predictive, analytic power of Big Data can affect everything from the Census and surveillance to citizen engagement and providing resources, thus becoming a central record of government activity.[40] InterPARES 2 found that VanMap, a GIS system for the City of Vancouver that aggregates government data for subsequent analysis and decision-making, "provides evidence of transactions . . . aggregates and presents information in ways that facilitate new activities and transactions" and "is both an instrument and a by-product of the practical activities of its creator."[41] Following InterPARES 2's conclusion that aspects of VanMap should be captured and treated as records, this article argues that other Big Data systems, particularly in government contexts, have the capacity to generate records and should be managed as such.

So far, this discussion has elucidated *why* Big Data should be preserved and its role in an archives. It is now time to consider *how* Big Data can be managed for successful long-term preservation. While an extended technical explanation of preserving Big Data is beyond the scope of this article, there are projects that provide guidance on what such a practice could look like. The aforementioned InterPARES 2 project, while examining the creation, maintenance, and long-term preservation of authentic experiential, interactive, and dynamic records in electronic systems,[42] introduced the concept of "bounded variability," which it defines as "the changes to the form and/or content of a digital record that are limited and controlled by fixed rules, so that the same query, request or interaction always generates the same result."[43] Because Big Data is so dynamic, in its current state, it likely does not meet the requirements for bounded variability. Unless Big Data is actively set aside upon its use, it will not have limits and fixed rules that allow for the recall of a set variation of results. However, following the findings of the VanMap case study, bounded variability can be built into a Big Data system to support its long-term preservation.[44] A system that captures snapshots of a Big Data set when it is modified could produce records. Apache Hadoop, a cloud computing platform

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

that supports Big Data processing, includes the Hadoop Distributed File System (HDFS), which can take snapshots of partial or entire file systems.[45] Researchers at the University of Applied Sciences Eastern Switzerland HTW Chur developed a prototype for a mixed mode data repository (MMRepo) using Hadoop 2.0 and HDFS2.[46] Ingo Barkow, Catharina Wasner, and Fabian Odoni do not explicitly address the HDFS2 snapshot capability, which normally requires user involvement for a snapshot to be taken. However, Tsozen Yeh and Yipin Wang sought to address this limitation by "revis[ing] and improv[ing] the snapshot scheme" so that "when changes [are] made to a part of the file system with snapshots taken before, [the] system will autonomously take a new snapshot on that part of the file system in real time."[47] Yeh and Wang conceptualize the automatic snapshot in terms of restoring backups after errors, rather than preserving authentic records. Interdisciplinary collaboration is needed to bridge technical developments with archival theory. Archivists have a key role to play in the design and maintenance of Big Data systems to ensure their recordkeeping capabilities.

One inherent component of Big Data that presents a particular challenge to digital preservation is size. Capturing iterations of a Big Data system requires significant technical and financial resources for storage alone. In the context of digital libraries, Wasim Ahmad Bhat argues that "the huge storage shortage for long-term preservation of big data can be largely attributed to the failure of the storage technology to cope with the growth of big data."[48] While storage is important, technological developments will likely address the issue of digital storage space. The real challenge created by the volume of Big Data is the complexity of access and use.[49] Because Big Data requires computational support to be understandable, users will want to interact with data sets in dynamic ways to make sense of the records. The complexity of sense-making may pose technical challenges, but archival interventions can mitigate them in part. Metadata creation and management as part of an effective digital preservation program can help contextualize the records, but also provide key information for future renderings of Big Data.

Proper metadata is essential for ensuring the authenticity of Big Data in the long run. Just as records are centered around provenance, so, too, is data lineage essential to future use and understanding.[50] The Digital Curation Centre's (DCC) Curation Lifecycle Model speaks to the importance of managing data from the moment of creation.[51] Big Data requires knowledge not only of how its current aggregations are functioning, but also of the data sets' initial origins. The DCC advises generating administrative, descriptive, structural, technical, and preservation metadata at the time of creation. Because digital preservation ideally begins at record creation, the arrangement and description of Big Data is intertwined with preservation actions. When preserving Big Data, a central question is: what is being preserved? Boyd and Crawford emphasize the importance of

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

taking a critical stance toward Big Data, rather than assuming objectivity. It is important to remember that "Big Data are heterogeneous collections, created in varied sites of production and shaped by their conflicting values and norms."[52] In addition to the technical requirements of preservation, it is essential that archivists capture the complex provenance of Big Data. Because Big Data involves the aggregation of data, it can combine data sets from a wide variety of contexts. Within governments, this can involve the integration of various departments' data sets to gain bigger picture understandings of services. "Big Data are assembled from local conditions that are important for understanding the whole . . . by looking at Big Data as agglomerations of local data, we can learn about the heterogeneity of data in general and the importance of data's origins."[53] From the perspective of long-term preservation and in accordance with the DCC model, this heterogeneity is part of Big Data's identity and thus must be properly documented to maintain the data's authenticity. Additionally, the multiprovenancial origins of Big Data may require specific work to ensure metadata interoperability across data sources.

Yanni Alexander Loukissas argues that recognizing heterogeneity is key to understanding the wider social implications of Big Data. Daniela Agostinho argues that "under the guise of neutrality, the claims to objective knowledge staked out by big data often try to elide the active role of data in shaping a world that is becoming increasingly datafiable."[54] While Agostinho is not speaking about professional archival practice or Big Data as records, her arguments about the power dynamics of information have implications for archives. Archivists cannot make claims to the neutrality or objectivity of Big Data, but they can use their theoretical knowledge to represent the reliability, or lack thereof, of Big Data records. Agostinho reflects on the "political implications" of "the collection and usage of big data" given that appraisal "always entails operations of exclusion . . . with the hyping of big data there is the heightened danger of excluding all data deemed small whilst privileging the potentially big."[55] This concern for the dynamic between "big" and "small" data has specific implications for government archives because government Big Data can be deeply personal, and its use can have immediate and significant impacts on the lives of individuals. The SAMI archives requires complex systems to manage galaxy data, but in the government context, an added layer of complexity arises from archives including data about people. Despite the large-scale nature of Big Data, it is important to remember that personal information, and the personal stakes invested in that information, are key components of these data sets. Without critical perspectives that acknowledge the implications of humans becoming data points, Big Data archives may reproduce systems of inequalities that are often intertwined with government surveillance and services.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Archivists can play a key role in changing the power dynamics of Big Data. Many authors discuss the challenges of understanding the context of Big Data, but there is often no connection to the role archives can play in representing and preserving context.[56] Boyd and Crawford argue, "Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality."[57] These questions about knowledge, organization, and understanding are important, but difficult to parse out with Big Data because it is not always clear what the data sources are. To make claims about data, "we need to know where data is coming from; it is similarly important to know and account for the weaknesses in that data."[58] Two issues are the ways in which Big Data can compound inaccuracies and the need for different methods of understanding Big Data. Regarding inaccuracies, boyd and Crawford note that "data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together."[59] While inaccurate content does not preclude Big Data from having record potential, lack of context, and thus lack of archival bond, does. Archivists nonetheless should be deeply aware of the potential for inaccuracies as essential context. If long-term preservers of Big Data do not account for inaccuracy, they risk inadvertently perpetuating information and knowledge production that is not only wrong, but potentially harmful. Understanding provenance is key here. Returning to boyd and Crawford's claim that Big Data has "the aura of truth, objectivity, and accuracy," archivists can use their skills to break down these mythologies by situating Big Data in its context of creation.[60] In archives, representing Big Data records as reliable, rather than objective, will help users understand this context. While record accuracy is part of record creation practices, ultimately, recordkeepers and archivists seek to keep authentic records, which can include capturing data with content errors. Contextualizing these errors is also important because of the potential for archives themselves to become sources of Big Data.[61] An extended discussion of this phenomenon is out of scope in this article, but, as technology continues to develop, archivists should be aware of the ways holdings will become subject to Big Data analytics.

Transparently representing Big Data inaccuracies is one method of mitigating the mythology of "truth, objectivity, and accuracy."[62] As boyd and Crawford note, these myths are part of Big Data's creation and use within its original activities. Because archives seek to preserve authentic records, rather than truthful or objective ones, they must represent these records accurately and critically. Context is key because Big Data is particularly difficult to understand, and, given its use across disciplines and sectors, it is becoming a significant underpinning of current research. Loukissas argues that "attention to the local is necessary for developing critical discourses on Big Data."[63] He warns against

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

the "the universalizing ambitions of Big Data," instead emphasizing the need to understand the heterogeneous origins of data sets.[64] Heterogeneity can itself be an area of research; Loukissas points to "the utility of differences in data as markers of an otherwise invisible local context that is important for meaningful analysis."[65] Beyond supporting diverse forms of research, understanding data provenance illuminates social power dynamics. "Data are situated within the means of their production, the infrastructure required to maintain them, their systems of representation, and the social order they reproduce."[66] The context of creation shows the activities of those who produce and use Big Data, especially in governments. Loukissas's argument also reminds us that these activities can partially depend on their technological context. The software and hardware required to aggregate and analyze data sets can be seen "as evidence of the way that data are situated in time."[67] Understanding how governments use Big Data technology to perform their day-to-day functional activities is another way to hold them accountable. Digital preservation, then, plays a key role in managing technological context for long-term access. The need to know this context underscores the reasons why Big Data must be preserved as systems; as previously discussed, bounded variability can be built into Big Data systems so there is the potential to preserve Big Data as dynamic records that serve as or support evidence of activities.

The wide impacts of Big Data can be serious and need to be captured when Big Data becomes records. Because of the dynamism and supposed anonymization of Big Data, citizens have "diminished access to our own data, we often lack the means and expertise to analyze, make sense of it, even recognize it as our own."[68] While Big Data can support open government, and initiatives geared toward accountability and transparency, it can also support opaque bureaucratic structures in which citizens have little or no control over their personal information. As previously noted, Big Data is not immune to inaccuracies, and, if data are inaccurate, once aggregated in a Big Data context, the inaccuracy becomes difficult to rectify. In Australia, an "apparent wrong-headed deployment of big data" by the government resulted in claims of people "being wrongfully threatened with legal action for failure to pay debts that do not exist."[69] This example shows the real-world effects of inaccurate Big Data, which will be compounded over time without proper documentation and contextualization. Providing this context is one of many ways that reliable recordkeeping can help larger societal goals. Authentic Big Data records may in fact evidence problems in a government's data collection that have serious implications for the treatment of its citizens. By preserving Big Data and providing access to citizens, the archives can be a space where individuals are re-empowered to engage with their data.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Providing access to Big Data, in combination with contextualizing it, is by no means simple, but this is another area in which archivists have useful expertise to contribute. Others worry that "without the software and hardware of [legacy data sets'] era, as well as operating knowledge thereof, data would not be accessible at all."[70] Big Data from legacy systems left unmanaged will become inaccessible, but intervention by archivists can make it accessible in the long term. Emulation may present one path forward, but still runs the risk of inaccessibility due to proprietary software rights and obsolescence.[71] Processes like the HDFS automatic snapshot in Hadoop indicate feasible options for capturing records that might then be managed in the long term outside of their original systems. Regardless of the specific strategies, archivists must play an active role in Big Data generation and management to successfully retain its accessibility. Original technological context is important and can be captured through emulation and/or description, but, ultimately, Big Data should be maintained and preserved in a way that encourages usability for a wide array of audiences independent of a requirement for specific legacy software and hardware. Considering widespread accessibility is important because "it is still necessary to ask critical questions about what all this data means, who gets access to what data, how data analysis is deployed, and to what ends."[72] In the government context, Big Data is used to provide services and regulate the population. Citizens should have a means for accessing this information. "Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access."[73] Access is, however, just the first step in making Big Data available for purposes of transparency and accountability. Returning to Agostinho's argument that "we often lack the means and expertise to analyze [Big Data], make sense of it, even recognize it as our own," Big Data compounds the problem of access by requiring more than literal availability to be understandable.[74] There are no simple solutions, but when considering the preservation of Big Data, it is important to recognize the additional work required to make data truly accessible to general users.

While the facilitation of access may be a means of citizen empowerment, it is also essential to be aware of the privacy issues embedded in Big Data records. Locating Big Data about oneself is challenging, but it is inaccurate to consider Big Data sets as anonymous. The Australian government was embroiled in more Big Data controversy when it decided to link the 2016 census with data from other agencies and to link individuals' census data into the future to track them across time. The Australian Bureau of Statistics planned on monetizing the data, which raises serious concerns about citizen privacy and consent.[75] Given the relationship between government data collection and the provision of essential services, citizens may not be able to opt out of data collection.[76] If governments then use Big Data methodologies, citizens have little control over

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

personal information. As Galloway argues, "once we have given our information, our privacy has already been breached."[77] Privacy breaches are an increased risk with Big Data because of the potential for "deductive disclosure," which is "facilitated by the volume of data and its complexity, with the result that it is difficult to fully assess beforehand the risk."[78] Deductive disclosure is a process whereby personal information is accessible through the combination of large, individually anonymized data sets, because the aggregation of that data generates connections between data that make individuals identifiable. Ultimately, Big Data raises the stakes in the relationship between people's private and public lives. Theoretically, anonymized data disconnects citizens from having agency over their personal information, yet the processes of data linking increasingly amalgamate all aspects of people making them more exposed. Through Big Data, governments can take macro-level perspectives of their citizens, yet it is important to remember that behind data points are human beings whose lives the information collected about them deeply affect. This personal aspect of Big Data only grows as long-term preservation becomes a factor; the accumulation of data sets also represents the accumulation of lived experiences and personal information that must be protected. When preserving government Big Data for long-term access and use, archivists must be aware of the increased risks regarding privacy breaches. Understanding the provenance of each data set and appropriately documenting their contexts of creation and use may help determine the potential risks. Similarly, intervening at earlier stages of system design can give archivists the opportunity to raise issues of privacy and access restrictions.

Archives exist not only to preserve records, but also to provide access to them. Given the serious privacy risks embedded in government Big Data, why preserve data at all? The appraisal of Big Data is a separate discussion, but as governments increasingly turn to Big Data to carry out their functions, it will inevitably become part of the modern government archives. While archivists must be aware of the privacy implications of Big Data, particularly in the context of mass government surveillance, it is also essential to be prepared to ingest these records into archival holdings. The volume, velocity, variety, and veracity of Big Data pose unique technical and ethical challenges to digital preservation. Taking a systems-level view of Big Data by attempting to capture instances of bounded variability may be one path forward, and technical tools and systems can successfully manage such large volumes of information. However, ultimately, as with all digital preservation initiatives, proper documentation is key. Following in the footsteps of current data curation projects, creating appropriate metadata to capture the identity, technical characteristics, and management actions for Big Data must include the multiprovenancial origins of such data sets. More broadly, Big Data reminds archivists of their larger responsibilities.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Recognizing the power dynamics in Big Data requires an interrogation and documentation of the data themselves, as well as of the ways governments and corporations use them. Digital preservation must balance technical knowledge with critical perspectives to truly capture the context of Big Data and the records it produces.

Notes

[1] danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon, *Information, Communication and Society* 15, no. 5 (2012): 663, https://doi.org/10.1080/1369118X.2012.678878.

[2] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: SAGE Publications, 2014), 68.

[3] Syed S Husain, Alexandr Kalinin, Anh Truong, and Ivo D Dinov, "SOCR Data Dashboard: An Integrated Big Data Archive Mashing Medicare, Labor, Census and Econometric Information," *Journal of Big Data* 2, no. 1 (2015): 2, https://doi.org/10.1186/s40537-015-0018-z.

[4] Kitchin, *The Data Revolution*, 68.

[5] IBM, Big Data & Analytics Hub, "The Four Vs of Big Data," https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[6] boyd and Crawford, "Critical Questions for Big Data," 663.

[7] boyd and Crawford, "Critical Questions for Big Data," 663.

[8] I. S. Konstantopoulos et al., "The SAMI Galaxy Survey: A Prototype Data Archive for Big Science Exploration," *Astronomy and Computing* 13 (2015): 58–66, https://doi.org/10.1016/j.ascom.2015.08.002.

[9] Kitchin, *The Data Revolution*; Melissa de Zwart, Sal Humphreys, and Beatrix van Dissel, "Surveillance, Big Data and Democracy: Lessons for Australia from the US and UK," *University of New South Wales Law Journal* 37, no. 2 (2014): 713–47.

[10] Kitchin, *The Data Revolution*, 114; de Zwart, Humphreys, and van Dissel, "Surveillance, Big Data and Democracy"; see also Fola Malomo and Vania Sena, "Data Intelligence for Local Government? Assessing the Benefits and Barriers to Use of Big Data in the Public Sector," *Policy and Internet* 9, no. 1 (2017): 7–27, https://doi.org/10.1002/poi3.141.

[11] Malomo and Sena, "Data Intelligence for Local Government?," 8.

[12] Malomo and Sena, "Data Intelligence for Local Government?," 9.

[13] Kate Galloway, "Big Data: A Case Study of Disruption and Government Power," *Alternative Law Journal* 42, no. 2 (2017): 89–95, https://doi.org/10.1177/1037969X17710612.

[14] Kitchin, *The Data Revolution*, 113.

[15] Kitchin, *The Data Revolution*, 116.

[16] Kitchin, *The Data Revolution*, 88.

[17] PRISM is a program of the US National Security Agency that collects communications from US internet companies.

[18] While President Barack Obama's campaigns were historic for several reasons, they were among the first to successfully leverage the power of Big Data to identify and reach potential voters. Kitchin, *The Data Revolution*, 75.

[19] Kitchin, *The Data Revolution*, 75, 116.

[20] boyd and Crawford, "Critical Questions for Big Data"; de Zwart, Humphreys, and van Dissel, "Surveillance, Big Data and Democracy"; Galloway, "Big Data."

[21] Husain et al., "SOCR Data Dashboard."

[22] SAMI is one of many ways of considering data curation within the sciences. Bicarregui et al. also examine the management of Big Data in Big Science contexts, while Chen and Chen investigate Big Data case studies related to the geosciences. See Juan Bicarregui et al., "Data Management and Preservation Planning for Big Science," *International Journal of Digital Curation* 8, no. 1 (2013):

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

29–41, https://doi.org/10.2218/ijdc.v8i1.247; Suzhen Chen and Bin Chen, "Practices, Challenges and Prospects of Big Data Curation: A Case Study in Geoscience," *Internal Journal of Digital Curation* 14, no. 1 (2020): 275–91, https://doi.org/10.22181ijdc.v14i1.669.

[23] Konstantopoulos et al., "The SAMI Galaxy Survey," 59–62.

[24] Cristina Ribeiro and Maria Eugénia Matos-Fernandes, "Data Curation at U. Porto: Identifying Current Practices across Disciplinary Domains," *IASSIST Quarterly* 35, no. 4 (2011): 14–17, https://doi.org/10.29173/iq893.

[25] Chen and Chen, "Practices, Challenges and Prospects of Big Data Curation," 279.

[26] Chen and Chen, "Practices, Challenges and Prospects of Big Data Curation," 278.

[27] Konstantopoulos et al., "The SAMI Galaxy Survey"; Chen and Chen, "Practices, Challenges and Prospects of Big Data Curation."

[28] Chen and Chen, "Practices, Challenges and Prospects of Big Data Curation," 283.

[29] Tracey P. Lauriault, Barbara L. Craig, and InterPARES 2 Project, "General Study 10 Final Report: Preservation Practices of Scientific Data Portals" (2008), 44.

[30] Limor Peer and Stephanie Wykstra, "New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation," *IASSIST Quarterly* 39, no. 4 (2015): 6–13, https://doi.org/10.29173/iq902; Ingo Barkow, Catharina Wasner, and Fabian Odoni, "MMRepo—Storing Qualitative and Quantitative Data into One Big Data Repository," *IASSIST Quarterly* 40, no. 4 (2016): 14–19, https://doi.org/10.29173/iq908; Alex H. Poole and Deborah A. Garwood, '"Natural Allies: Librarians, Archivists, and Big Data in International Digital Humanities Project Work," *Journal of Documentation* 74, no. 4 (2018): 804–26, https://doi.org/10.1108/JD-10-2017-0137.

[31] Poole and Garwood, "'Natural Allies: Librarians, Archivists, and Big Data in International Digital Humanities Project Work."

[32] InterPARES 2 Terminology Database, http://web.archive.org/web/20200214223401/http://www.interpares.org/ip2/ip2_terminology_db.cfm.

[33] Defined by the InterPARES 2 Terminology Database as "an indivisible unit of information constituted by a message affixed to a medium (recorded) in a stable syntactic manner. A document has fixed form and stable content."

[34] Yvette Hackett, "Part Four—Methods of Appraisal and Preservation: Domain 3 Task Force Report," in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, ed. Luciana Duranti and Randy Preston, vol. 19 (Padova: Associazione Nazionale Archivistica Italiana, 2008), 29, https://doi.org/10.1108/rmj.2009.28119aae.003.

[35] Luciana Duranti and Kenneth Thibodeau, "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES," *Archival Science* 6, no. 1 (2006): 13–68, https://doi.org/10.1007/s10502-006-9021-7.

[36] The InterPARES 2 Terminology Database defines an *interactive record* as "a record with variable content or form that is dependent on user input that is often based on earlier content."

[37] The InterPARES 2 Terminology Database defines a *dynamic record* as "a record the content of which is dependent upon data that might have variable instantiations and be held in databases and spreadsheets internal or external to the system in which the record is generated."

[38] Galloway, "Big Data."

[39] Chen and Chen, "Practices, Challenges and Prospects of Big Data Curation," 276.

[40] Galloway, "Big Data"; Malomo and Sena, "Data Intelligence for Local Government?"; Kitchin, *The Data Revolution*.

[41] Glenn Dingwall, Richard Marciano, Reagan Moore, and Evelyn Peters McLellan, "From Data to Records: Preserving the Geographic Information System of the City of Vancouver," *Archivaria* 64 (2007): 181–98.

[42] InterPARES 2 Project: Project Summary, http://web.archive.org/web/20191002193415/http://www.interpares.org/ip2/ip2_index.cfm

[43] InterPARES 2 Terminology Database.

[44] Dingwall et al., "From Data to Records."

[45] Tsozen Yeh and Yipin Wang, "Enhancing Hadoop System Dependability through Autonomous Snapshot," *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on*

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (2018), 653–60, https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00116.

[46] Barkow, Wasner, and Odoni, "MMRepo—Storing Qualitative and Quantitative Data into One Big Data Repository."

[47] Yeh and Wang, "Enhancing Hadoop System Dependability through Autonomous Snapshot," 653.

[48] Wasim Ahmad Bhat, "Long-Term Preservation of Big Data: Prospects of Current Storage Technologies in Digital Libraries," *Library Hi Tech* 36, no. 3 (2018): 539–55, https://doi.org/10.1108/LHT-06-2017-0117.

[49] Konstantopoulos et al., "The SAMI Galaxy Survey," 58–59.

[50] Lauriault, Craig, and InterPARES 2 Project, "General Study 10 Final Report: Preservation Practices of Scientific Data Portals."

[51] Digital Curation Centre, "DCC Curation Lifecycle Model," http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[52] Yanni Alexander Loukissas, "Taking Big Data Apart: Local Readings of Composite Media Collections," *Information, Communication & Society* 20, no. 5 (2017): 652, https://doi.org/10.1080/1369118X.2016.1211722.

[53] Loukissas, "Taking Big Data Apart," 653.

[54] Daniela Agostinho, "Big Data, Time and the Archive," *Symplok* 24, nos. 1–2 (2016): 441, https://doi.org/10.5250/symploke.24.1-2.0435.

[55] Agostinho, "Big Data, Time and the Archive," 440.

[56] For example, see Loukissas, "Taking Big Data Apart"; boyd and Crawford, "Critical Questions for Big Data"; de Zwart, Humphreys, and van Dissel, "Surveillance, Big Data and Democracy."

[57] boyd and Crawford, "Critical Questions for Big Data," 665.

[58] boyd and Crawford, "Critical Questions for Big Data," 668.

[59] boyd and Crawford, "Critical Questions for Big Data," 668.

[60] boyd and Crawford, "Critical Questions for Big Data," 663.

[61] Devon Mordell, "Critical Question for Archives as (Big) Data," *Archivaria* 87 (2019): 140–61.

[62] boyd and Crawford, "Critical Questions for Big Data," 663.

[63] Loukissas, "Taking Big Data Apart," 652.

[64] Loukissas, "Taking Big Data Apart," 652.

[65] Loukissas, "Taking Big Data Apart," 657.

[66] Loukissas, "Taking Big Data Apart," 656.

[67] Loukissas, "Taking Big Data Apart," 655.

[68] Agostinho, "Big Data, Time and the Archive," 441.

[69] Galloway, "Big Data," 93. Centrelink, Australia's social security payment department, matched its data with the Australian Taxation Office to discover inconsistencies in claims. Because of how the data were matched, it produced a high number of inconsistencies even when claims were valid.

[70] Loukissas, "Taking Big Data Apart," 655.

[71] Sarah Mason, "Introduction to Digital Preservation: Emulation," *Oxford LibGuides* 28 (August 2018), https://libguides.bodleian.ox.ac.uk/digitalpreservation/emulation.

[72] boyd and Crawford, "Critical Questions for Big Data," 664.

[73] boyd and Crawford, "Critical Questions for Big Data," 674.

[74] Agostinho, "Big Data, Time and the Archive," 441.

[75] Galloway, "Big Data," 91–92.

[76] de Zwart, Humphreys, and van Dissel, "Surveillance, Big Data and Democracy."

[77] Galloway, "Big Data," 92.

[78] Malomo and Sena, "Data Intelligence for Local Government?," 14.

Emily Larson is a digital systems consultant at the UBC Residential School History and Dialogue Centre, where she focuses on user experience, interactive and emerging technologies, and processing.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.

ABOUT THE AUTHOR

She has an MAS and MLIS from the University of British Columbia School of Information. As an information professional, her work centers on the importance of storytelling and the power dynamics of information. She is located in Vancouver, Canada, on the unceded territories of the Musqueam, Squamish, and Tsleil-Waututh Nations. She is the recipient of the 2019 Theodore Calvin Pease Award from the Society of American Archivists (SAA). The award recognizes superior writing achievements by students of archival studies and was presented on August 4, 2019, during the SAA Annual Meeting. Larson's paper, "Big Questions: Digital Preservation of Big Data in Government" was nominated by Luciana Duranti, professor of archival studies at the University of British Columbia in Vancouver.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-29 via Open Access.