Big Data Curation Framework: Curation Actions and Challenges

Ayoung Yoon, Jihyun Kim, and Devan Ray Donaldson

Abstract
Big data curation represents an emerging topic of inquiry but still in an early phase along its adoption curve. The term big data itself is a nebulous concept and the differences between small data curation and big data curation are nuanced. The goal of this research is to provide a theoretical framework that identifies big data curation actions and associated curation challenges. This study is based on the practices of big data research and data curation by systematically examining literature. The outcome of the study includes the big data curation framework that provides overview of curation activities as well as concerns that are essential to perform such activities. The study also provides practical implications for libraries, archives, data repositories, and other information organizations that concerns the issue of big data curation as big data presents a multi-dimensional array of exigencies in relation to the mission of those organizations.

Keywords
Big data, big data curation, data curation, framework development

1. Introduction

Generally defined as "high-volume, high-velocity and/or high-variety information assets" [1], big data has emerged as the basis of modern research in many areas, such as medicine, environmental science, and urban planning [2]. This positions big data stewardship to be a vital part of scientific research by addressing big data storage, processing, and use [2]. Traditionally, libraries have been responsible for curating data as part of their mission to protect and disseminate data [3]. Beginning in the mid-2000s, literature addressed the role of libraries in managing what was initially dubbed the "data deluge" (e.g., [4]). Hey and Hey [5] announced that libraries would have to leverage capabilities developed in managing assets, such as institutional repositories and scholarly communication infrastructure, metadata, and information retrieval systems, to meet the demand for research data management services created by the data deluge. Blummer and Kenton [4] also argued that librarians' unique skills (e.g., collection development, cataloging, preservation, and searching) and experiences with user studies to meet users' needs and assessment qualified them for big data curation. By using or translating librarians' skills and knowledge, libraries need to play a role in curating and providing access to big data [6].

While academic libraries have been more involved in data curation through research data management services suited for data-intensive applications and services [7, 8], public libraries are also slowly recognizing the need to be involved in data curation due to the rise in publicly available data. Many open government data initiatives at the federal, state, and city levels currently exist or are in the process of being created, and these initiatives often involve partnerships with public libraries. Huwe [9] discussed how a public library can work with local actors to provide data access and sharing to promote better governance. Thus, librarians need to embrace their role in making big data more useful, visible, and accessible by creating taxonomies, designing metadata schemes, and systematizing retrieval methods [10]. Many have asserted the necessity of building sustainable and scalable data curation programs in libraries to

This is the author's manuscript of the article published in final edited form as:

Yoon, A., Kim, J., & Donaldson, D. R. (2022). Big data curation framework: Curation actions and challenges. Journal of Information Science, 01655515221133528. https://doi.org/10.1177/01655515221133528

effectively continue their support of communities and scholarship that may be overwhelmed by big data [11].

While the importance of supporting big data research through proper big data curation services is clear, the meaning of big data curation has not been well-defined. Challenges with curating big data were first discussed around the mid-1990s, as data that were too large to fit local computer memory started to become an issue. While recent technological developments have made significant progress in handling data volume, researchers still argue that not all challenges of big data, such as variety and velocity, are addressed by traditional data curation efforts [12]. Because big data are fundamentally different from "spreadsheet science," curating them requires "innovative forms of processing that enable enhanced insight, decision-making and process automation" [1]. However, only a few articles have directly attempted to describe big data curation (e.g., [6, 12, 13]), despite that efforts to understand data characteristics relevant to big data have been continuous (e.g., creating data curation taxonomy; [14, 15]). Possible reasons for this include that the term "big data" itself is a nebulous concept. Likewise, the unique characteristics of big data curation are nuanced and therefore not within the scope of data curation research. A frequent assertion in the literature is that small data curation is the most effective approach to improving data quality for big data [16]. Finally, big data curation represents an emerging topic of inquiry that is still in its early phases [13].

Our study aims to provide a theoretical framework that identifies big data curation actions and associated curation challenges in the context of big data. The purposes of our literature review were to identify high-level attributes of big data curation from the previous literature and to understand the challenges associated with big data research and practices that impact big data curation. Our study is based on not only the theoretical discussion on big data curation actions and the lifecycle but also the practices of big data researchers who have concerned with and/or utilized big data. Our study contributes to the domain of libraries, archives, data repositories, and other information organizations concerned with the issue of big data curation. Despite agreement on the importance of having big data curation programs in this domain, little previous effort has been made to systematically understand the concept of big data curation in such organizations. Our framework can provide a foundation for many organizations wanting to develop big data curation programs, as it provides an overview of curation activities and concerns, which is essential to identifying the necessary resources to develop such a program. While our framework is function-based, as it is the first step toward identifying big data curation actions, we plan to further expand the framework with a role-oriented approach in our future study, mapping those curation actions with actors who will conduct those activities.

2. Literature Review

Originating in the mid-1990s at Silicon Graphics [17], the term "big data" became widespread by the 2000s [18]. Gandomi and Haider [18] argued that the current hype can be attributed to the promotional initiatives by IBM and other leading technology companies that invested in building the niche analytics market. While the term "big data" itself has been widely used, the definition of big data has not been well-defined. It is most commonly characterized by its challenges, typically referred to as the 4Vs: volume, variety, velocity, and veracity [19, 20]. Volume refers to the size and scale of data, variety to the heterogeneity and different forms of data, velocity to the

rate at which data arrive and the time frame in which they must be acted upon, and veracity to the quality [12, 19, 21, 22]. Emphasizing the socio-cultural aspects of big data, other researchers have added additional characteristics to ascribe to big data, such as value [21] and usability and privacy [19]. Kitchin [23] added exhaustive in scope, fine-grained resolution, relational, flexible, and scalable to his big data definition.

Perhaps one of the most important considerations when defining big data is that the "big" does not describe size but rather the "capacity to search, aggregate, and cross-reference large data sets" [24, p. 663]. Jacobs [25, p. 39] argued that "the truth about big data in traditional databases [is] it's easier to get the data in than out" for analysis. Thus, big data has less to do with size and more to do with what is required to analyze the data as well as what tools are available.

Big data curation is even less well-defined than big data. It is commonly referred to as "scalable" data curation that responds to the challenges of big data, such as "data volume, variety, etc., thereby increasing the value of maintaining [a] large dataset" [13, pp. 87-88). Big data curation is also seen as the application of the 4Vs to the data lifecycle [12], thereby adding value to large datasets, increasing their discoverability and accessibility, and ensuring their long-term preservation. Other researchers have emphasized that the nature of big data addresses new computational research methods and cyberinfrastructure tied in with big data analytics [26]. According to Kune et al. [27, p.95], a large part of big data curation should include the "'Middleware' layer of [the] big data cloud architecture model".

Big data presents a multidimensional array of exigencies in relation to the mission of libraries, archives, and repositories. Lynch [28] emphasized the duty of libraries to provide data management because of the precarious nature of valuable scientific data, which run the risk of being lost if they do not receive the proper care. While data curation corresponds to key library activities(e.g., collections development; metadata management; instruction, liaison, and outreach; provision of access to resources; and infrastructure), big data multiplies the complexity of these core activities. How should data curation services for research data obtained using big data techniques proceed? Due to the expansion of cloud services and distributed computing with high-level interfaces, big data is not restricted to major institutions that are able to maintain the necessary infrastructure for high-performance computing; big data tools and techniques, parallel and concurrent data processing, and virtualization have become commonplace. To cope with this shifting paradigm, big data curation must entail reworking some established LIS models of practice, such as OAIS [29, 30].

Existing studies have discussed the risks of investing in big data, including the high cost of infrastructure and expertise and the uncertain value of the results. In the library context, some libraries that invested too heavily in data science activities failed to succeed in this highly competitive environment [31]. Another risk is vendor lock-in as well as the more general threat of losing authority and identity—and thereby compromising some of the ethical underpinnings of librarianship—due to the partnerships necessary to deliver big data services [10, 32, 33]. Despite these risks, Teets and Goldner [6] presented one of the earliest arguments for big data curation as its own domain and one in which the LIS profession should be involved, particularly through the deployment of linked data and semantic ontologies.

A few other scholars have addressed the issue of big data curation, arguing for the urgency of developing a more solid framework as well as proposing different models. Freitas and Curry [13] identified big data curation as originating in the academic domain with the sciences, but they predicted that the economic demand for data quality would lead to the expansion of the data curator role. They contended that "as an emergent activity, there is still vagueness and poor understanding on the role of data curation inside the big data lifecycle" [13, p. 7].While the costs of data curation are often not estimated during project planning, their point of focus was the "individuation and recognition of the *data curator role*, "which depends upon a realistic assessment of "the costs associated with producing high-quality data" [13, p. 7]. Xie et al. [30] made a similar assertion about the need for shared cyberinfrastructure to handle big data in scholarly digital repositories by having a direct role in the big data lifecycle and using cloud computing and containerization to provide access to big data analytics.

While those studies provided meaningful groundwork for studying big data curation, few researchers have identified detailed curation actions and associated curation challenges, which must be done first before defining the role of curators. Pouchard [12] proposed the framework of the data lifecycle model as the device by which big data curation can be mapped and understood as a domain, but most previous works have been based on theoretical discussions rather than empirical big data research and practices. Our study fills the gaps in the existing research by systematically examining big data research and the mapped associated curation actions.

3. Methods

To develop a big data curation framework, the project team conducted a systematic literature review to search the relevant literature and identify big data curation actions. The systematic review was the best approach for our project because it is a structured method that can identify and critically analyze a vast amount of research outcomes in a research field. Systematic reviews establish to what extent the existing research has progressed to clarify a particular problem and are useful for identifying gaps, relations, and any contradictions in existing studies. Moreover, systematic reviews are useful for proposing a new conceptualization or theory and particularly for formulating an overarching conceptualization to comment on, extend, or develop a new theory [34, 35, 36].

For our study goal, we adopted and slightly modified the five steps of conducting a systematic review originally proposed by Khan, Kunz, Kleijnen, and Antes [37]:

Step 1: Framing the questions for a review,
Step 2: Identifying relevant work,
Step 3: Assessing the quality of studies,
Step 4: Summarizing the findings, and
Step 5: Interpreting the findings.

Step 1

We framed the following questions based on our topic of big data curation: What actions relevant to big data curation have been used by researchers in big data projects (empirical evidence from behaviors)? What key big data curation actions have been suggested by data

curation professionals or researchers (theoretical evidence from conceptual discussions)? Our goal was not only to review articles that directly discussed big data curation models (theoretical) but also those that actually conducted big data research that revealed one or more curation-associated actions during the research cycle (empirical). As already noted, only a few existing studies have directly discussed big data curation models or activities. Further, many studies that have utilized big data have presented issues or problems associated with big data curation, even though those studies either did not focus on discussing curation issues or were not aware of the relationship with big data curation. By including literature that not only directly discussed the concept of big data curation but also studies that utilized big data and thus had implications for big data curation, our big data curation framework is based on both theoretical and empirical evidence.

Step 2

To identify relevant studies, we employed inclusion or exclusion criteria, such as terms, language, restrictions, and other limits, as suggested by McKibbon [38]. Our inclusion criteria were journal articles, conference proceedings, and book chapters published after 2012 in order to focus on recent research. Additionally, we only included works written in English. We excluded letters, editorials, book reviews, research notes, and short communications as well as articles published in other languages.

We utilized five different search terms: "big data curation," "big data management," "big data challenges," "big data problems," and "big data issues." Because the term "big data curation" is not well-adopted in all disciplines that may utilize big data research, the term "big data management" was also employed because it is often used when addressing data curation issues. We also utilized "big data challenges," "big data problems," and "big data issues" because these terms are related to data curation issues and challenges. We conducted a keyword search in three major databases: Web of Science, ScienceDirect, and EBSCOhost. We also conducted additional searches in Google Scholar as well as major information and library science journals, information technology journals, and conference proceedings (e.g., *Journal of the Association of Information Science and Technology*, *International Journal of Digital Curation*, *Library and Information Science Research*, *Library Hi-Tech*, and *IEEE International Conference on Big Data*). Our search was conducted in three phases: 1) database searches were conducted from September to October 2017, 2) LIS-specific searchers were conducted from January to February 2018, and 3) additional searches for works published in 2018 and 2019 were conducted in summer 2020 while the project team was analyzing the data. From those series of searches, we collected 492 articles from the database search and 112 articles from the LIS-specific search, for a total of 604 articles.

Step 3

In order to assess the quality of the literature for our study, we used Zotero bibliography management software. We first created a folder from each search result, imported the search results, attached PDFs for all articles, and manually assessed each article for its quality and how well it fit our purposes based on the abstract. Irrelevant (e.g., not containing any aspect of data curation), duplicate (e.g., from more than one search result), and low-quality (e.g., simply

summarized other works) articles were removed during this process. As we needed to closely examine the articles for their quality, an iterative process was used and continued during Step 4 (data analysis). Once we finalized the list of articles, a total of 177 articles remained, 118 of which were from the database search and 59 from the LIS-specific search.

Step 4

We analyzed the articles identified from Step 3 using a pre-developed protocol. The protocol included a control number for each article, originating fields/disciplines of the articles, article types (theoretical research employing research methods, empirical research, conceptual essay without research methods, and literature review), definitions, significance, challenges, and pre-defined data curation/management actions. We adopted the big data lifecycle model proposed by Pouchard [12] with modifications to pre-define data curation actions in our protocol. Pouchard's [12] model is theoretical and data oriented, presenting big data research practices with project-level curation concerns. This model was useful for this study, as it is one of few models to address the characteristics of big data (4Vs), highlighting the potential impact on data management and curation associated with those characteristics. We also adopted the research data curation framework proposed by the Educause Working Group [39], which is a group that consists of various data curation experts (e.g., librarians, researchers, and IT experts). This framework was adopted for its service-oriented approach to addressing resources, collaboration, partnership, and ethics and policies beyond the project level. We reviewed the common steps and considerations in these two models to create our research framework addressing the data lifecycle from the service perspective (see Table 1). In our framework, the asterisk indicates iterative processes throughout the curation procedure. The obelisk indicates actions that may not be part of the data curation process, such as a research process, but that may still offer some important considerations for curation actions.

Table 1. Pre-developed protocol for big data curation action analysis

Model	**Actions**							
Big data lifecycle model [12]	Plan	Acquire	Describe*	Prepare	Analyze	Assure*	Preserve	Discover
Research data curation framework [39]	Curation planning	-	Description	Preparation and analysis		-	Storage	Discovery/ access
Big data curation framework (our analysis model)	Planning	Acquisition†	Description*	Preparation	Analysis†	Assurance*	Storage and preservation	Discovery/ access

Asterisks (*) indicate iterative process throughout the cycle
Obelisk (†)indicates the processes that may offer considerations for curation actions.

Based on our understanding of data curation, the project team thoroughly examined 177 articles and captured the relevant activities that were associated with big data and the curation lifecycle,

even if the articles reviewed did not specifically identify or discuss curation activities. Additionally, we coded for big data curation challenges discussed in those articles. Inter-coder reliability was 90.1%, which was an acceptable rate. Once the coding was completed, we invited an external validator (second author of this paper) to review the coding quality.

Step 5

Once we finished our analysis and identified big data curation actions, we proceeded with the interpretation by distinguishing curation actions particular to each characteristic of big data (4Vs), mapped with the curation challenges identified from Step 4. From this, we created a table of curation actions per the 4Vs and their associated challenges, which was our initial big data curation framework. Then, we identified the gaps in existing big data curation actions, discussed the important considerations for big data curation missed in our literature analysis, added those missing components to the chart, and created our final big data curation framework (see Table 2).

4. Results

4.1.Summary of the Literature

The majority of the articles were published during the late2010s (24 in 2014, 39 each in 2015 and 2016, and 44 in 2017). About half of the papers were theoretical research (23%) or literature reviews (17%), but there werealso a good number of empirical research papers(17%). In terms of the disciplinary distribution, about 28% of articles were from computer science disciplines (according to the journal/conference proceedings categorization), but a wide variety of other disciplines were also identified, such as business, engineering, LIS, geography, economics, earth science, medicine, archaeology, architecture, astronomy, psychology, education, and sociology, which represent a growing interest in and the relevance of big data in those disciplines.

4.2.Big Data Curation Actions and Considerations

Planning

Planning is the first step and makes curation activities more efficient throughout the curation lifecycle [40]. A total of 48 articles discussed actions relevant to planning. First, setting a vision and strategy for big data curation is an underlying condition for successful planning to guide the overall process [41]. Such an approach will encourage organizations to think more about how big data curation aligns with their overall missions, core values, and service models. Having big data curation as part of the organizational strategy will also create incentives and help generate economic models for data curation [13, 42]. Second, it is important to recognize, identify, and define the roles and responsibilities of curators, relevant stakeholders, and different departments within and/or across organizations [13, 41, 43]. Such a process encompasses investigating the inclusion of internal and external resources as well as different vendors and products [44]. Understanding and assigning curation roles will eventually facilitate long-term thinking about future strategizing, resource and risk assessment, and cost analysis [42, 43, 44].

It is also important to define data stages, ideally from the lifecycle analysis [45], and to determine the types of data the curators will handle and preserve and the necessary tools [12, 46]. Data governance issues should be considered at all stages, as this will allow for the specification of decisions about rights, accountability, ownership, and accessibility [47, 48]. A governance mechanism may be created from cross-department coordination efforts, and such initiatives are essential to producing agreement on any data processing and further curation actions [41].

Acquisition

Acquisition is the data collection process based on the pre-set criteria to produce, generate, and ingest data for the research process. As explained in Table 1, acquisition may not be a necessary part of data curation actions. However, the nature of the data collected during this process as well as the methods used may influence overall curation actions, such as description, preparation, and assurance. Thus, it is still important to understand the process of acquisition. A total of 42 articles discussed the actions relevant to acquisition with considerations for curation actions. For instance, the importance of understanding the nature of data during source identification (e.g., how the data is collected and identifying datasets that need to be combined from multiple data sources) was discussed in several studies (e.g., [13, 49, 50]). Data conditions, their heterogeneous nature, and the relationship between different sources directly contribute to the complexity of the data structure and quality. Such conditions may require more processing in later stages of the lifecycle [51, 52, 53, 54]. Therefore, a data examination should be conducted during the acquisition and selection process for future curation actions such as description, assurance, and preservation. Preprocessing may occur simultaneously alongside data crawling of each data source (e.g., extraction of the date, topic, and relation) to ensure that the data are properly collected and represented for the remaining process [55, 56].

During data acquisition, researchers can decide on the level of data management actions needed depending on the type of data being collected, as some data are highly structured (e.g., sensor data), whereas others (e.g., web data) are highly heterogeneous [48]. Several studies argued for the need to think about the database model (e.g., NoSQL, cloud, event-based, spatiotemporal, or relational) at this stage to ensure the reshaping of heterogeneous data and to provide a unified data operation in later stages [56, 57, 58]. Finally, given the massive variety of data, massive ingestion may be necessary, which sometimes involves crowdsourcing, scheduled crawling, or automation [59, 60, 61, 62]. Nonetheless, the precedent should be the metadata standards for different data types upon which diverse communities agree [61].

Description

Description involves creating, collecting, preserving, and maintaining sufficient metadata [63]. A total of 44 articles discussed aspects of description. It is necessary to define a large set of metadata that describe a big data lifecycle, use appropriate tools for automated capture of the metadata, and develop metadata management practices [50]. Metadata identify what data are created and how they are measured [64]. Additionally, metadata provide provenance information about data processing [9, 65]; associated knowledge such as mission, equipment, data structure, or format [43]; data version management information [66, 67, 68]; and persistent identifiers [48, 69].

Using standardized metadata schemes, vocabularies, and ontologies is essential during the description stage to address the varieties of syntaxes, semantics, and formats, thereby increasing interoperability [13, 70, 71]. However, it is challenging to connect different schemes or data models from diverse disciplines [72]. To deal with data variety, Tardio et al. [45] suggested using algorithms to calculate similarities among different data models. Semantic standards, such as RDF or OWL, are used to describe and integrate big data [41, 73]. Linked data representation can be used to integrate different data and make them usable on the web [6, 52, 74].

Pouchard [12] suggested that scalable tools that enable automatic generation and extraction of metadata are needed for describing big data. Adopting semantic and automated scientific workflow tools makes it possible to immediately capture provenance and work history metadata. Park and Brenza [75] noted the advantages of semi-automatically created metadata in terms of scalability, cost-effectiveness, and consistency. Automatic metadata generation techniques were also identified, including meta-tag harvesting, content extraction, automatic indexing, text and data mining, extrinsic data auto generation, and social tagging [76]. In particular, indexing methods or software applicable for big data were suggested to help users interpret and utilize data in a flexible and effective manner [49, 62, 77].

Preparation

A total of 38 articles in this study discussed the importance of and actions relevant to preparation. Data preparation is essential for future data use and storage [56], but its complexity is often overlooked, and the processes can be very time consuming [12]. Freitas and Curry [13] argued for the need for data to be systematically processed, transformed, and repurposed into a new context in order to extract the value of the data.

Examples of actions relevant to data preparation include data cleaning (e.g., pre-processing, such as filling in missing values or mapping attributes) and transformation (e.g., normalization, standardization, and data integration/aggregation with metamodels) [44, 50, 78, 79]. Data cleaning and pre-processing is an important first step toward removing noisy and redundant data; this is followed by data reduction and compression [44, 65, 79]. Data conversion and transformation is also an important step in making data more user friendly and machine readable [43], which enables data sharing and fusion between heterogeneous data sources [80]. Data integration is vital to data preparation, as it adds value to datasets taken from multiple sources of raw data [45, 81]. Still, merging data from a diversity of sources of differing quality is challenging [53], which is why developing a multidimensional data model is necessary [45]. Data transformation needs to be pre-conducted, as different data formats need corresponding data models for standardized and unified forms [82].

A few considerations emerged regarding data preparation. First, anonymization of sensitive data is important [83]. Second, because big data management requires a great deal of manual preparation and intervention (e.g., decoding databases and assigning data types manually), automation of this process based on semantic technology is necessary [52, 84]. Third, to support efficient data preparation, researchers need to use appropriate technical infrastructure, such as cloud computing [40]; a distributed big data platform, such as the Hadoop with NoSQL database

[85]; real-time data processing technology to find errors, conflicts, and missing data [86]; and a new data model that can negotiate between relational and non-relational database systems [48].

Assurance

Researchers need to be assured of not only the data quality but also their provenance and usage for security, privacy, and trust. A total of 33 articles addressed activities relevant to assurance. Assessing the quality of big data is important because quality directly relates to veracity [12], which indicates the data's accuracy [86, 87, 88], completeness [12, 73, 89], and uncertainty [88]. Maintaining data consistency improves their quality [52, 73]. Specific methods for enhancing big data quality have been suggested in terms of developing an iterative methodology via the continuous integration of new and existing data [45], standardizing image protocols [47], and using metrics to evaluate the data's value [50].

Tracking the provenance of big data is essential to ensuring the trustworthiness of data [13, 90] as well as their privacy and security [91]. It is necessary to develop tools for managing data provenance to establish data-level trust and enable permission management [13, 92, 93]. The literature frequently stated concerns about privacy and confidentiality (e.g., [9, 53, 94]. Lawrence [95] also noted that separating storage into long-term archives versus active science is necessary for security.

Analysis

The analysis phase of the big data lifecycle involves the use of various tools and analytical methods to make informed decisions [12]. While analysis itself is not part of data curation activity, this process still has several implications for curation, as big data analysis involves different methods, tools, hardware, and underlying architecture. Data visualization tools can be used to present comprehensive information in a dynamic and interactive manner [29, 55, 58, 70, 81, 85, 96]. For example, the Hadoop-based MapReduce big data processing mechanism is used to extract insights from data [73, 90, 97]. Additionally, data summarization tools are needed to create knowledge from complex data [52]. The tools used for data mining, statistical analysis, and visualization can be helpful for curators [51, 69], enabling them to engage in complex data analysis [98] and allowing for scientific reproducibility [12]. Researchers can also choose to preserve software and source codes used during analysis as well as the description of their features, the toolkit versions used in producing the code, and the range of meaningful values for input parameters [12]. Because these objects are not traditionally part of data curation, and the underlying architectures age quickly, preserving them can present new challenges [12].

Storage and Preservation

In total, 30 articles discussed actions taken during their big data research. Most of these were concerned with storage issues regarding real-time data access and transferring data from the data processing unit to a storage space for preservation [99]. Big data researchers commonly discussed two layers of architecture: hardware infrastructure (e.g., high-performance computing equipped with volume and velocity) and management infrastructure (e.g., software with an interface that allowed for quick retrieval of large-scale datasets) [100]. The use of different

storage architectures is necessary because of the heterogeneous nature of data (structured, unstructured, and semi-structured) [79]. Distributed storage systems, including Hadoop, Spark, and nonrelational database systems (e.g., HBase and MongoDB), were commonly used [59, 62, 80]. The cloud platform was also frequently used as a way to gather data from and share them with multiple locations [54, 82].

Big data researchers also mentioned several actions from previous lifecycles, such as assigning metadata and documenting data processes, which are essential for future use. Standardizing the metadata and format was also discussed, as this step can also be done during storage and preservation [41, 43]. Because simply preserving data and metadata is insufficient for future reuse, curation of associated knowledge should also occur and be established during the planning and preparation stages [101, 102]. Finally, big data storage infrastructures should ensure data security (e.g., integrity and confidentiality) and the protection of data ownership [12].

Discovery and Access

The discovery and access step is a core component of big data curation, as it is critical to data reuse, making it possible to uncover patterns and insights from the existing data [81]. Information professionals, whose reference and resource discovery skills remain valid in big data practices, play an important role in making data discoverable and accessible by developing taxonomies, metadata schemes, and retrieval methods [9, 10]. A total of 48 articles discussed actions relevant to discovery and access. The main aspects of discovery and access were repositories or platforms that make the data accessible, standardized metadata, and ontologies that improve searchability [12, 103].

Disciplinary, institutional, and community-based repositories are important infrastructural features for big data discovery because they enhance access to and reliance upon data and metadata standards, federated search systems, software tools, persistent identifiers, ontologies, and interoperability supports [2, 12, 103, 104, 105]. Platforms are important for big data accessibility, and several studies specifically argued for the benefits of cloud computing in the big data context as it enables easy access to data through simplified and centralized control (e.g., [47, 92]). Cloud-based data can also be a cost-effective option for keeping up with current systems and new technologies, although this applies more to certain data types (e.g., GIS) [106]. This is achievable due to the use of low-end equipment and open-source software that address some problems relating to distributed storage, processing, querying, interoperability, and the virtualization of massive spatial data [58, 80].

While standardized metadata facilitate data discovery and access in any kind of data curation work, it is especially important in big data curation given the characteristics of big data (e.g., 4Vs). Standardizing query languages for each variety of big data structure is important in metadata retrieval [50]. Several studies explored different methods to enhance searchability by adopting standardized metadata formats and uniform schemas to improve interoperability across distributed repositories (e.g., [41, 43, 58, 97]). Still, existing strategies for improving query performance are not transferable to the context of big data [90], which requires new approaches, such as those that are time or location based, for improvement [52].

4.3. Challenges of Big Data Curation

While the analysis of the big data curation steps presented some challenges associated with each stage, several other major and overarching challenges to big data curation emerged.

Scalability

Scalability is a central challenge, since the traditional curation model could be potentially difficult to scale up to satisfy storage and service requirements [62, 107, 108]. Freitas and Curry [13] noted that addressing scalability is a multidisciplinary problem that requires the development of economic models, social structures, incentive models, and standards in coordination with technological solutions.

Storage Capacity

Due to the limitations of media, many researchers argued for the need for storage solutions that can support advanced analysis; interactive explorations; and particularly, a real-time query of large dynamic datasets [109, 110] with highly scalable file systems [29]. Deciding what to store seems to be another concern due to time consumption, cost, and the inefficiency of formally backing up peta scale data, but there exists no simple approach to identifying important data [95].

System Capacity

Many scholars also addressed the challenges to the system due to capacity issues, such as network bandwidth capacity for transferring, storing, and distributing data, in addition to concerns about system power and the imbalance of CPU-heavy (but I/O-poor) database systems that are not suitable for big data curation [111, 112]. Current data architecture is not prepared to deal with heterogeneous data [113], and standard systems may also lack the ability to support large amounts of data scattered across many different information systems, even once they are integrated into the centralized information system [62, 84]. The capability to respond in real time—or at least within an acceptable timeframe—was mentioned as a significant challenge by multiple big data researchers (e.g., [86, 114, 115, 116]).

Description and Standardization

Difficulties in standardization are apparent in big data curation. There is a lack of standards for data and data formats [10, 117, 118, 119], which exacerbates the issue of the heterogeneity of data [120]. To effectively share and analyze big data, it is important to provide data in a standardized manner with standard APIs and web protocols [6, 121]. Providing high-quality and descriptive metadata is needed for metadata standardization [121, 122]. Describing the provenance of the metadata (e.g., the scientific workflow whereby data are created, used, and modified) is also essential [68, 71, 123]. Clarifying terminologies, tools and algorithms, data characteristics, and scientific queries [123]—as well as developing a shared understanding of interpretations and actions—can lead to greater interoperability [124].

Human Resources

Beyond the technical challenges to supporting big data curation, there have been difficulties in allocating the appropriate levels of human resources and capital necessary to work with a large amount of data, allowing for the transformation of raw data into meaningful information [118]. Freitas and Curry [13] stated that there is still vagueness surrounding the role of data curation inside the big data lifecycle, which has resulted in underestimating data curation costs for many projects. Big data curation requires a high level of expertise in data management and distributed systems and databases, but there are gaps in resources and training for researchers and curators [69, 125, 126]. With rapid changes in technology and data growth, training and maintaining the knowledgebase's accuracy and currency seems to be an enormous challenge [126].

Privacy and Security

Privacy, security, and confidentiality concerns emerged strongly when multiple agencies, each with a different purpose, could access data [52]. There was special concern regarding the inevitability of cloud storage becoming a more common practice in the future and the security of this storage method, such as how the encryption of sensitive information can be maintained when the cloud is not 100% trusted and how data retrieval can be adapted to keep information safe [127, 128, 129]. Anonymization of sensitive information may lead to more challenges, as confidential information could be inadvertently transmitted to unintended parties as a result [113, 119].

5. Discussion

Our analysis presents big data curation actions related not only to theoretical research on the topic but also to real-world work practices utilizing big data in other research contexts. Our research included the perspectives and practices of various disciplines, ranging from engineering, astronomy, and computer science to education and sociology. Based on our analysis, we proposed a big data curation framework, presented in Table 2. Table 2 identifies the big data curation actions and their associated challenges, both from the literature analysis (normal font) and our interpretation of the findings that identified the gaps (*italic font*). The first column presents our analysis framework, which was developed during our analysis (see Methods section). The next five columns show curation actions identified from our analysis. The overview includes general actions that match each stage of the lifecycle, followed by actions specifically addressing bigdata characteristics. The last column and row summarize the challenges associated with either specific stages of the curation lifecycle or specific characteristics of big data.

Because our analysis model was adopted from existing bigdata curation models and theories, it is not unusual that general data curation principles and models are still applicable to bigdata curation practices. However, it is worth mentioning that there are big data-specific actions and challenges when it comes to practical implementation. Many such challenges are inherent to bigdata characteristics, particularly the 4Vs. With decentralized data creation in real time, the unprecedented scale of data brings foundational challenges that need to be managed. Furthermore, the nature of different datasets in terms of form, format, size, and embodied differences affects the different stages of the lifecycle, including data sharing, archiving,

Table 2. Big Data Curation Framework

Big Data Curation Framework	Curation Actions					Associated Challenges
	Overview	**Action Addressing Volume**	**Action Addressing Variety**	**Action Addressing Velocity**	**Action Addressing Veracity**	
Definition of 4Vs	-	Size and scale of data	Heterogeneity and different forms of data (*Most challenging, relevant to quality)	Rate at which data arrive and the time frame in which they must be acted upon	Quality attributes, such as biases, noise, and abnormalities	-
Planning	• Set a vision and strategy • Create incentives • Identify roles and responsibilities of curators and relevant stakeholders • Pre-set criteria for producing, collecting, and ingesting data	• Plan for infrastructure based on the volume • Assess need for increased data volume and increased demand for services	• *Integrate policies for different data types* • *Create policies to govern submission, usage, and services (1)*	*Accommodate based on input rate (2)*	• *Consider quality and ethics for overall curation workflow (3)* • Create data governance mechanism	• Incentive structures for curation needed • Existing services (e.g., libraries) must adapt to changes in data practices by developing new skills and infrastructure
Acquisition†	• Select, acquire, and ingest (raw data harvest): selection as "value-laden" decision • *Negotiate for closed data* • *Consult on data-sharing agreement* • Examine data for selection • Preprocess data • Prepare new data model to deal with 4Vs	Ingest on massive scale, including crowdsourcing, scheduled crawling, or automation	• Assessing data condition, heterogeneity, formats, types, and structure • Maintain semantic consistency	Determine frequency of data acquisition based on input rate	• *Consider the nature of sources and data in order to account for possible biases (4)* • *Obtain consent (5)*	*Curators' knowledge of research process, data analytics, and associated technology/tools*
Description*	• Document through automated semantic tools and/or data extraction • Crowdsource for annotation and description	Gather metadata and perform documentation at scale	• Examine different models of knowledge representation • Translate as needed to make	• Make decisions about levels of description/ curation	Track provenance in order to cope with unexpected problems and outcomes	• Data standardization • Data interoperability • Data provenance

	• Establish semantic tags that are backed up by standardized ontologies • Create pipelines and workflows to track dependencies between processes and data • Make connections to the raw data		disciplinary and interdisciplinary data manageable and understandable	• Automatically generate and extract metadata		
Preparation	• Transform and clean data • Integrate and aggregate data • Annotate data to create scaffold • Convert data to machine-readable files • Create appropriate technical infrastructure addressing 4Vs	• Transform and clean data at scale • Perform automation	• Clean to make heterogeneous data usable • Create workflows for different types of data • Design new data model/database system	• Utilize real-time data processing technology • *Select data for preservation based on input rate (6)*	Anonymize data	*Curator's knowledge and skills in data modeling and analytics*
Assurance*	• Validate data • Flag missing and/or incomplete data • Ensure privacy and security • *Perform document format conversion (7*)	Validate at scale: closely related to veracity and assurance of veracity (trust level)	Validate formats for various types of data	-	• Protect privacy of people who generated the data used for analysis • *Determine legal and ethical considerations (8)*	• Assuring data quality for users while also protecting sensitive data • Quality of open and protected data in a cloud computing environment
Analysis†	• Utilize statistical techniques to assess data quality (repetitive and overlapping data) • Model data • Mine data	Ensure analyzability at scale	• Transform data into less complex forms • Accommodate for emergent properties of data	-	• Deal with noise • Make code available for determining/supporting transparency	• Preservation of source code • Keeping up with hardware and underlying architecture
Storage & Preservation	• Identify repository and storage system	• Determine storage capacity • Determine hardware	Find methods for preserving a high variety of data (e.g., unstructured,	Transfer real-time data for preservation	-	Security of cloud storage

	• *Perform migration and source format management (proprietary vs. open source) (9)* • *Make preservation decisions (e.g., source code, raw data) (10)* • Build multi-tier storage architecture • Create distributed storage system • Standardize metadata and formats for storage and preservation	infrastructure to cope with volume and velocity	new media-based file format)			
Discovery & Access	• *Determine access for different levels of data (e.g., raw, processed, software code, end-outcome) (11)* • *Collaborate with researchers for resource discovery (12)* • *Provide "middleware" service rather than end-user activity (13)*	-	Standardize metadata and query languages	Utilize cloud-based repository platform	• Create policy for data sharing, privacy, and confidentiality • Perform information audits to prevent data abuse	• Libraries' participation in the resource discovery part of the data life cycle is insufficient • Indexing multidimensional data for object-based search and retrieval
Associated Challenges	*Human resources and capital (e.g., curation skills, project management, and directing participants in data curation projects)*	• Scalability of traditional curation model and computing power • Flexible and scalable infrastructures needed to ingest high-volume, heterogeneous datasets	• Lack of data standards and formats • Heterogeneity: machines cannot match depth of natural language • Inconsistency/in completeness problems with distributed data sources, crowdsourcing	• *Real-time technology to address all aspects of curation* • *Velocity can impact overall curation cycle by making the process iterative*	Privacy and confidentiality issues	-

* Drives all stages of data life cycle; metadata is key to maintaining interoperability

† Processes that may offer considerations for curation actions

discovery, organization and description, linkage, and interoperability, all of which will eventually impact the services provided by libraries and related infrastructures [72]. Understanding how each characteristic influences curation actions and their associated challenges is important in this regard. Our framework therefore clarifies the specific links between each bigdata characteristic and curation action as well as the associated challenges.

We also recognize that there are a few areas that were not covered in the literature analysis but are still important when curating big data. We added missing actions from the analysis, which are presented in italics with assigned numbers, to the framework. First, the literature analyzed in our study seemed to be less concerned with the social aspects of curation issues, especially policy and ethical considerations. The issue of privacy protection and security was mentioned mostly in relation to the assurance curation action. However, the implications of legal and ethical considerations should be broader, including the issues of data ownership, licenses, intellectual property, appropriate data storage, and ethical data sharing and reuse; thus, we added those components to our framework (see [8] in Table 2). Data governance policy should be developed, and the legal requirements of different types of data should be checked during the planning stage of the curation cycle (see [1] and [3] in Table 2). Sensitivity should also be an important consideration when selecting and acquiring data (see [4] and [5] in Table 2) because machine-learning algorithms can absorb unconscious biases (e.g., racism) in a population and have a tendency to help overrepresented populations and even possibly harm underrepresented populations [130]. Obtaining informed consent for certain types of big data (e.g., social media data) is not practical or desirable given the size and speed of data creation, but data curation should consider the idea that the lack of consent merely makes data more sensitive while not necessarily precluding the sharing of it [131].

Second, as presented in our results, the majority of the literature we examined was concerned with storage issues rather than preservation. Due to the lack of a few key preservation considerations, we added these to the framework (see [7], [9] and [10] in Table 2). These include migration and format management (proprietary vs. open source) and decisions about the levels of preservation (e.g., raw data, source code, and bit-level preservation). Preservation decisions should also be included in the planning stage, particularly regarding the selection of data for preservation based on both the data's value and input rate (see [6] in Table 2). Despite the plummeting cost of hard disk storage and open-source infrastructures like Hadoop, many bigdata stakeholders struggle to decide what data they should use, preserve, or dispose of. Additionally, capturing and storing all data in perpetuity may not be feasible forever.

Third, the literature review does not cover some aspects of access to support current or future reuse, such as access to different levels of data, collaboration with researchers to facilitate discovery, and possible middleware services to reduce the complexities of managing individual applications and systems. Thus, we added these to the framework (see [11], [12], and [13] in Table 2). These factors imply the need for specific tools and services that allow for flexible and efficient approaches to reusing big data. Providing access to big data, in combination with contextualizing them, is not simple, but it is the most critical role of data curation. Larson [132] argued that the complexity of access and (re)use is the real challenge. Using big data requires computational support to be understandable, and user interaction is dynamic; both require human

and technical interventions to mitigate the process. Ultimately, big data should be supported by independent usability without the requirement of specific legacy software and hardware, but there is no simple answer to how that should be done.

Our framework presents bigdata curation challenges from multidimensional perspectives considering the curation stages and the characteristics of big data. The range of challenges discussed includes the issues of scalability, system and computing capacity, description and standardization, human resources, and privacy and security. One notable aspect of bigdata curation challenges is human resources. For example, it is important to ensure that people are trained in curation skills and are able to manage curation projects and that there is adequate staff supervision in data curation projects—a factor that can be easily overlooked when focusing on the technical solutions of bigdata curation. While scaling curation involves technical solutions, such as automation and crowdsourcing, humans still need to intervene to direct algorithmic behavior, as reflected in the statement by Stephanie McReynolds, VP of Marketing at Alation: "Curations are about where the humans can actually add their knowledge to what the machine has automated" [133]. We also observed that there have been efforts to address several data curation challenges by developing new approaches to data models, testing new automation techniques, and designing new storage systems. However, those efforts happen in a silo, perhaps because many data curation needs are specific to certain data and project contexts (e.g., GIS and temporal data), and curating data often requires ad hoc solutions specific to the use case [134]. While there is no one-size-fits-all solution in data curation, it might still be necessary to move from small-scale, project-dependent approaches to a holistic understanding in order to address bigdata curation problems.

While our framework is the first effort of its kind to systematically understand the actions and challenges of big data curation, one limitation of this framework is that it does not lay out the relationships and roles of potential contributors and stakeholders in curation actions. Data curation is a collaborative process, and inter-departmental or inter-organizational collaboration is necessary to identify and bring all resources to the table, determine any missing pieces, and establish a workflow with clear roles for each contributor. Because our framework is action oriented, the relationships among relevant stakeholders as well as the desired resources to support each action are not integrated. While outside the scope of this research, mapping the desired resources, as well as the necessary skills and manpower, with identified curation actions would be useful for practitioners who want to interpret the framework and implement any bigdata curation plans.

6. Conclusion

Providing open access to research outcomes (e.g., publications, datasets, and computer codes) has become more common in many disciplines as the practice of open science is adopted to improve the reproducibility of science and original research. Incorporating open science practices into big data research often poses challenges due to the size of datasets, multilayer data architecture, computing powers and cloud systems, and the complexity of data sources. While data curation is key to supporting data sharing and open access, there are no standard solutions yet to address those challenges in a big data context. Lowndes et al. [135] emphasized the significance of extending existing scientific data management support, as big data acquisition,

collection, and processing activities can overwhelm a project, making it difficult to contribute open and reproducible science. While a number of data management and curation guidelines and tools exist, Dietrich et al. [136] noted that many lacks the details and a common set of standards to help adoption and implementation, especially in the context of big data.

The primary contribution of this study is that it bridges the gap in big data curation frameworks by providing a theoretical framework based on conceptual arguments along with the practice of bigdata curation in various fields. The framework balances conceptual and practical knowledge from the literature and helps connect abstract ideas to real-world applications. While this is a theoretical framework, it can be widely used to understand feasible aspects of big data curation and enhance its execution across diverse practical settings, particularly in the field of research computing, data services, and libraries, where a strategic approach is necessary to provide big data curation services when technology is rapidly changing, organizational structure and operation should be considered, and collaboration with campus IT is necessary.

The framework reinforces the scope of libraries' roles while suggesting that libraries should augment their traditional roles of preserving and ensuring access to human knowledge in bigdata contexts. Goldberg et al. [106] argued that because one of the primary roles of libraries in our society is to be the repository of human knowledge, libraries can also be a repository of data for long-term curation, which leverages a library's core strengths to make data discoverable and usable by the largest possible audience. In addition to long-term preservation, libraries seeking to determine their role in the handling of big data within their organization and to use it to develop big data services and accessibility is not a new concept [137]. By using internal skills to deal with data, libraries can support the analysis of big data to ensure data quality and help researchers and the public utilize these data to generate novel ideas and to shape their usefulness. Because of those potentials, several studies have discussed the possible roles of academic libraries [10, 51]; special libraries, like map and GIS data libraries [106]; medical libraries [138]; and public libraries [139] in big data curation. When implemented, these roles may expand their scope of work and allow for the scaling up of existing services.

One important area in which libraries can contribute to the landscape of big data curation is the formation of strategic partnerships with internal or external organizations to build robust infrastructure and tools that meet the needs of stakeholders. Previous studies pointed out unbalanced resource allocations for big data curation and use (e.g., [139]) and the lack of infrastructure to support access to and use of big data (e.g., [69, 92]) as the challenges of big data curation. These may pertain to libraries as well, such as technical barriers that prevent libraries from gaining access to appropriate support, technical connections, and capital investments [82]. However, libraries are skilled at building internal and external partnerships, including working with local communities and developers. While our framework does not discuss relationships among stakeholders or collaborations among relevant entities, it is clear that big data curation is not a one-person job due to its dynamic nature, distributed activities, and required domain expertise. Libraries can be the first place to bring different pieces of the puzzle—information technology, preservation expertise, research skills, and scientific knowledge—together to complete the picture, avoid silo work, and focus on the sharing of resources.

Acknowledgement

This research has funded by the Institute of Museum and Library Services (#LG-72-17-0139-17).

Reference

1. Gartner IT Glossary. Big data. http://www.gartner.com/it-glossary/big-data (accessed 21 March 2021).
2. Reinhalter L, Wittmann RJ. The library: Big data's boomtown. *Ser libr* 2014;67(4):363–372. https://doi.org/10.1080/0361526X.2014.915605
3. Heidorn PB. The emerging role of libraries in data curation and e-science. *J libr adm* 2011;51(7–8):662–672. https://doi.org/10.1080/01930826.2011.601269
4. Blummer B, Kenton JM. Big data and libraries: Identifying themes in the literature. *Internet Reference Services Quarterly* 2019 Jan 12;23(1-2):15–40. https://doi.org/10.1080/10875301.2018.1524337
5. Hey T, Hey J. e-Science and its implications for the library community. *Library Hi Tech* 2006;24(4):515-528. https://doi.org/10.1108/07378830610715383
6. Teets M, Goldner M. Libraries' role in curating and exposing big data. *Future Internet* 2013;5(3):429–438. https://doi.org/10.3390/fi5030429
7. Lyon L. The informatics transform: Re-engineering libraries for the data decade. *Int. J. Digit. Curation* 2012;7(1):126–138. https://doi.org/10.2218/ijdc.v7i1.220
8. Martinez-Uribe L, Macdonald S. A new role for the academic librarian: Data curation. *Profesionel de La Informacion* 2008;17(3):273–280. https://doi.org/10.3145/epi.2008.may.03
9. Huwe TK. Data discovery and data curation going hand in hand. *Comput libr* 2013;33(3):17-19.
10. Bieraugel M. Keeping up with... big data. *Association of College & Research Libraries (ACRL)*, http://www.ala.org/acrl/publications/keeping_up_with/big_data (2013, accessed 21 March 2021).
11. NISO. Research data curation, Part 2: Libraries and big data. http://www.niso.org/news/events/2013/webinars/data_curation/ (2013, accessed 9 May 2020).
12. Pouchard L. Revisiting the data lifecycle with big data curation. *Int. J. Digit. Curation* 2015;10(2):176–192. https://doi.org/10.2218/ijdc.v10i2.342
13. Freitas A, Curry E. Chapter 6 Big Data Curation. In: Cavanillas JM, Curry E and Wahlster W (eds) *New Horizons for a Data-Driven Economy A Roadmap for Usage and Exploitation of Big Data in Europe*. 1st ed. Springer International Publishing, 2016, pp. 87–118.
14. Chao TC, Cragin MH, Palmer CL. Data practices and curation vocabulary (DPCVocab): An empirically derived framework of scientific data practices and Curatorial Processes. *J Assoc Inf Sci Technol* 9 May 2014;66(3):616–33. https://doi.org/10.1002/asi.23184
15. Siddiqa A, Hashem I, Gani A, et al. A survey of big data management: Taxonomy and state-of-the-art. *J. Netw. Comput. Appl.* 2016;71: 151-166. https://doi.org/10.1016/j.jnca.2016.04.008
16. Morgan A, Duffield N, Hall LW. Research data management support: Sharing our experiences. *J. Aust. Libr. Inf. Assoc.* 2017;66(3):299–305. https://doi.org/10.1080/24750158.2017.1371911
17. Diebold, F. "Big Data" and its origins. https://doi.org/10.48550/arXiv.2008.05835 (2020, accessed 9 May 2020).
18. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Int J Inf Manage.* 2015;35(2):137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
19. Jagadish HV, Gehrke J, Labrinidis A, et al. Big data and its technical challenges. *Commun ACM*. 2014;57(7):86–94. https://doi.org/10.1145/2611567
20. Laney D. 3D data management: Controlling data volume, velocity, and variety (Application delivery strategies). *Meta Group, Inc*. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (2001, accessed 9 May 2020).
21. Demchencko Y, Grosso P, Laat C, et al. Addressing big data issues in Scientific Data Infrastructure. 2013 *Intl Conf Collab Tech Syst (CTS)*. 2013; 48-55, doi: 10.1109/CTS.2013.6567203.
22. IBM. The four V's of big data. *IBM Big Data & Analytics Hub*. http://www.ibmbigdatahub.com/infographic/four-vs-big-data (n.d., accessed 20 Jan 2018).

23. Kitchin R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc* 2014:1(1). https://doi.org/10.1177/2053951714528481
24. Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.*10 May 2012;15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878
25. Jacobs A. The pathologies of big data. *Commun ACM*. 2009;52(8):36–44. https://doi.org/10.1145/1536616.1536632
26. Witt M. Institutional repositories and research data curation in a distributed environment. *Libr trends* 2008;57(2):191–201.
27. Kune R, Konugurthi PK, Agarwal A, et al. The anatomy of big data computing. *Software Prac Exper* 2016;46(1):79–105. https://doi.org/10.1002/spe.2374
28. Lynch C. Big data: How do your data grow? *Nature* 2008;455(7209):28–29. https://doi.org/10.1038/455028a
29. Gerrard DM, Mooney JE, Thompson D. Digital preservation at big data scales: Proposing a step-change in preservation system architectures. *Libr. Hi Tech* 2018;36(3):524–538. https://doi.org/10.1108/LHT-06-2017-0122
30. Xie Z, Chen Y, Jiang T, et al. On-demand big data analysis in digital repositories. In: *Digital Libraries: Providing Quality Information. ICADL 2015* (ed R Allen, J Hunter, M Zeng), 2015. Springer International Publishing. https://doi.org/10.1007/978-3-319-27974-9_29
31. Huwe TK. Librarians and data: Curator, creator, or both? *Online Searcher* May/June 2017;41(3):10–15.
32. Palatin J, Lagoze C, Edwards P, et al. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media Soc* 2018;20(1): 293–310. https://doi.org/10.1177/1461444816661553
33. Padilla T. Collections as data: Implications for enclosure. *Coll res libr news* 2018;79(6):296–300. https://doi.org/10.5860/crln.79.6.296
34. Baumeister RF, Leary MR. Writing narrative literature reviews. *Rev. Gen. Psychol.* September 1997;1(3):311–320. https://doi.org/10.1037/1089-2680.1.3.3
35. Bem DJ. Writing a review article for Psychological Bulletin*. Psychol bull*. 1995;118(2):172–177. https://doi.org/10.1037/0033-2909.118.2.172
36. Cooper H. Editorial. *Psychol bull*. 2003;129:3–9. https://doi.org/10.1037/0033-2909.129.1.3
37. Khan K, Kunz R, Kleijnen J, et al. Five steps to conducting a systematic review. *J R Soc Med* 2013;96(3):118–121. http://doi.org/10.1258/jrsm.96.3.118
38. McKibbon A. Systematic reviews and librarians. *Libr trends* 2006;55(1):202-215. https://doi.org/10.1353/lib.2006.0049
39. Choudhury S, Cowles E, Croft HR, et al. Research data curation: A framework for an institution-wide services approach. EDUCAUSE Working Group on Data Curation May 2018. https://hsrc.himmelfarb.gwu.edu/libfacpubs/35
40. Ouf S, Nasr M. Cloud computing: The future of big data management. *Int. J. Cloud Comput.* 2015;5(2):53–61. https://doi.org/10.4018/IJCAC.2015040104
41. Ng ST, Xu FJ, Yang Y, et al. A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning. In*: Creative Construction Conference 2017*, Primosten, Croatia, 19-22 June 2017. pp. 939–947.
42. Frankova P, Drahogova M, Balco P. Agile project management approach and its use in big data management. *Procedia Comput Sci* 2016;83:576–583.
43. Kiemle S, Molch K, Schropp S, et al. Big data management in Earth observation: The German satellite data archive at the German Aerospace Center. *IEEE Geoscience and Remote Sensing Magazine*, September 2016;4(3):51–58. https://doi.org/10.1109/MGRS.2016.2541306
44. Orosz T, Orosz I. Company level big data management. In: *2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI),* Timisoara, Romania, 15-17 May 2014. pp. 299–303. https://doi.org/10.1109/SACI.2014.6840081

45. Tardio R, Mate A, Trujillo J. An iterative methodology for big data management, analysis, and visualization. In: *2015 IEEE International Conference on Big Data*. CA, USA, 28 December 2015. p. 545 -550. https://doi.org/10.1109/BigData.2015.7363798.
46. Stantic B, Pokornỳ J. Opportunities in big data management and processing. In: HM Haav, A Kalja and T Robal (eds) *Databases and Information Systems VIII*. IOS Press, 2014, pp. 15–26.
47. Brink JA. Big data management, access, and protection. *J Am Coll Radiol* 1 May 2017;14(5):579–580. https://doi.org/10.1016/j.jacr.2017.03.024
48. Kanchi S, Sandilya S, Ramkrishna S, et al. Challenges and solutions in big data management – An Overview. In: *2015 3rd International Conference on Future Internet of Things and Cloud*, Rome, Italy, 24-26 August 2015, pp. 418–426. https://doi.org/10.1109/FiCloud.2015.121
49. Abe A. Curating and mining (big) data. In: *2013 IEEE 13th International Conference on Data Mining Workshops*, TX, USA, 1-10 December 2013, pp. 664-671. https://doi.org/10.1109/ICDMW.2013.51
50. Storey VC, Song IY. Big data technologies and management: What conceptual modeling can do. *Data Knowl Eng*. 2017;108:50–67. https://doi.org/10.1016/j.datak.2017.01.001
51. Al-Barashdi H, Al-Karousi, R. Big Data in academic libraries: Literature review and future research directions. *J. Inf. Technol. (JIS&T)* 9 January 2019; 2018(2). https://doi.org/10.5339/jist.2018.13
52. Barnaghi P, Sheth A, Henson C. From data to actionable knowledge: Big data challenges in the web of things [Guest Editors' Introduction]. *IEEE Intell Syst* 2013;28(6):6–11. https://doi.org/10.1109/MIS.2013.142
53. Cyganek B, Grana M, Krawczyk B, et al. A survey of big data issues in electronic health record analysis. *Appl Artif Intell* 21 July 2016;30(6):497–520. https://doi.org/10.1080/08839514.2016.1193714
54. Wang H, Xu Z, Pedrycz W. An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges, and opportunities. *Knowl Based Syst* 15 February 2017;118:15–30. https://doi.org/10.1016/j.knosys.2016.11.008
55. Choi S, Seo J, Kim M, et al. Chronological big data curation: A study on the enhanced information retrieval system. *IEEE Access* 21 December 2016 ;5:11269–11277. https://doi.org/10.1109/ACCESS.2016.2642979
56. Hassani A, Gahnouchi SA. A framework for business process data management based on big data approach. *Procedia Comput Sci*, 2017;121:740–747. https://doi.org/10.1016/j.procs.2017.11.096
57. Lee M, Zhang Y, Chen S, et al. Heuristics for assessing Computational Archival Science (CAS) research: The case of the human face of big data project. In: *2017 IEEE International Conference on Big Data (Big Data)*, MA, USA, 11-14 December 2017, pp. 2262–2270. https://doi.org/10.1109/BigData.2017.8258179
58. Li W, Wu S, Song M, et al. A scalable cyberinfrastructure solution to support big data management and multivariate visualization of time-series sensor observation data. *Earth Sci Inf* 2016;9(4):449-464. https://doi.org/10.1007/s12145-016-0267-1
59. Cox R, Shah S, Frederick W, et al. A case study in creating transparency in using cultural big data: The legacy of slavery project. In: *2018 IEEE International Conference on Big Data (Big Data)*.WA, USA, 10-13 December 2018, p. 2689–2695. IEEE. https://doi.org/10.1109/BigData.2018.8621932
60. Liu J, Wang X, Khattak AJ, et al. How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis. *Neurocomputing* 2016;181:38–52. https://doi.org/10.1016/j.neucom.2015.08.098
61. Pierce L. Big data issues for remote sensing: Variety. In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, China, 10-15 July 2016, pp. 7593–7596. IEEE. https://doi.org/10.1109/IGARSS.2016.7730980
62. Wu Z, Wu J, Khabsa M, et al. Towards building a scholarly big data platform: Challenges, lessons and opportunities. In: *IEEE/ACM Joint Conference on Digital Libraries*, London, UK, 8-12 September 2014, pp. 117–126. IEEE. https://doi.org/10.1109/JCDL.2014.6970157
63. Ball A. *Review of the state of the art of the digital curation of research data*. Bath: University of Bath; 2010. http://opus.bath.ac.uk/19022/

64. Shao Y, Liu R, Wang F, et al. Research on big data management for high-speed railway equipment. *Appl. Mech. Mater.* 2014;462–463:405-409. https://doi.org/10.4028/www.scientific.net/amm.462-463.405
65. Demchenko Y, Grosso P, Laat C, et al. Addressing big data issues in scientific data infrastructure. In: *2013 International Conference on Collaboration Technologies and Systems (CTS),* CA, USA, 20-24 May 2013, pp. 48–55. https://doi.org/10.1109/CTS.2013.6567203
66. Gudivada VN, Rao D, Raghavan VV. NoSQL Systems for Big Data Management. In: *2014 IEEE World Congress on Services*, AK, USA, 27 Jun–2 July 2014, pp. 190–197. IEEE. https://doi.org/10.1109/SERVICES.2014.42
67. Johnson V. Leveraging technical library expertise for big data management. *Aust. Libr. Inf. Assoc.* 2017;66:271-286. https://doi.org/10.1080/24750158.2017.1356982
68. Macduff M, Lee B, Beus S. Versioning complex data. In: *2014 IEEE International Congress on Big Data*, AK, USA, 27 June-2 July 2014, pp. 788–791. IEEE. https://doi.org/10.1109/BigData.Congress.2014.124
69. Federer L. Research data management in the age of big data: Roles and opportunities for librarians. *Inf Serv Use* 2016;36(1–2):35–43. https://doi.org/10.3233/ISU-160797
70. Matsubayashi M, Kurata K. Conceptual design for comprehensive research support platform: Successful research data management generating big data from little data. In: *2017 IEEE International Conference on Big Data,* MA, USA, 11-14 December 2017, pp. 4407–4409. IEEE. https://doi.org/10.1109/BigData.2017.8258475
71. Padhy S, Jansen G, Alameda J, et al. Brown dog: Leveraging everything towards autocuration. In: *2015 IEEE International Conference on Big Data (Big Data)*, CA, USA, 29 October-1 November 2015, p. 493–500. IEEE. https://doi.org/10.1109/BigData.2015.7363791
72. Golub K, Hansson J. (Big) data in library and information science: a brief overview of some important problem areas. *J Univers. Comput Sci.* 2017;23(11):1098–1108. https://doi.org/10.3217/jucs-023-11-1098
73. Mountasser I, Ouhbi B, Frikh B. From data to wisdom: A new multi-layer prototype for Big Data management process. In: *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, Marrakech, Morocco, 14-16 December 2015, pp. 104–109. IEEE. https://doi.org/10.1109/ISDA.2015.7489209
74. Harper LM, Oltmann SM. Big data's impact on privacy for librarians and information professionals. *Bull Assoc Inf Sci Technol*April/May 2017;43(4):19–23. https://doi.org/10.1002/bul2.2017.1720430406
75. Park J, Brenza A. Evaluation of semi-automatic metadata generation tools: A Survey of the current state of the art. *Inf Technol Libr* 2015;34(3):22–42. https://doi.org/10.6017/ital.v34i3.5889
76. Polfreman M, Broughton V, Wilson A. Metadata generation for resource discovery. *JISC*. https://textarchive.ru/c-2308841-pall.html (2007, accessed 10 March 2021).
77. Lu W, Chen X, Ho DCW, et al. Analysis of the construction waste management performance in Hong Kong: The public and private sectors compared using big data. *J Cleaner Prod* 2016;112:521–531. https://doi.org/10.1016/j.jclepro.2015.06.106
78. Heer J, Shneiderman B. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue* 2012;10(2):30-55. https://doi.org/10.1145/2133416.2146416
79. Zhang Y, Ren S, Liu Y, et al. A framework for Big Data driven product lifecycle management. *J Cleaner Prod* 2017;159:229-240. https://doi.org/10.1016/j.jclepro.2017.04.172
80. Zhu Y, Tan Y, Luo X, et al. Big data management for cloud-enabled geological information services. *Sci Program* 2018;2018:1–13. https://doi.org/10.1155/2018/1327214
81. Zhou K, Fu C, Yang S. Big data driven smart energy management: From big data to big insights. *Renewable Sustainable Energy Rev* 2016;56:215–225. https://doi.org/10.1016/j.rser.2015.11.050
82. Liu Y. Research on the application of big data in academic libraries. In*: 2018 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, Xiamen, China, 25-26 January 2018, pp. 364–367. https://doi.org/10.1109/ICITBS.2018.00099

83. Sarjapur K, Suma V, Christa S, et al. Big data management system for personal privacy using SW and SDF. In: Satapathy S, Mandal J, Udgata S, Bhateja V (eds) *Information Systems Design and Intelligent Applications,* New Delhi: Springer, 2016, pp. 757–763. https://doi.org/10.1007/978-81-322-2752-6_75
84. Eine B, Jurisch M, Quint W. Ontology-based big data management. *Systems* 2017;5(3):45. https://doi.org/10.3390/systems5030045
85. Kim M, Choi J, Yoon J. Development of the big data management system on National Virtual Power Plant. In: *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, Krakow, Poland, 4-6 November 2015, pp.100–107. https://doi.org/10.1109/3PGCIC.2015.101
86. Yang C, Puthal D, Mohanty SP, et al. Big-sensing-data curation for the cloud is coming: A promise of scalable cloud-data-center mitigation for next-generation IOT and wireless sensor networks. *IEEE Consum Electron Mag*. 22 September 2017;6(4):48–56. https://doi.org/10.1109/MCE.2017.2714695
87. Gui H, Zheng R, Ma C, et al. An architecture for healthcare big data management and analysis. In: Yin X, Geller J, Li Y, Zhou R, Wang H, Zhang Y (eds) *Health Information Science*. Springer International Publishing, 2016, pp. 154–160. https://doi.org/10.1007/978-3-319-48335-1_17
88. Cuzzocrea A. Temporal aspects of big data management: State-of-the-art analysis and future research directions. In: *2015 22nd International Symposium on Temporal Representation and Reasoning (TIME)*, Kassel, Germany, 23-25 September 2015, p. 180–185. IEEE. https://doi.org/10.1109/TIME.2015.31
89. Keimle S, Molch K, Schropp S, et al. Big Data Management in Earth Observation: The German satellite data archive at the German Aerospace Center. *IEEE Trans. Geosci. Remote Sens.* 2016;4(3): 51-58. doi: 10.1109/MGRS.2016.2541306.
90. Eltabakh MY. Data organization and curation in big data. In: Zomaya, AY, Sakr S, (eds) *Handbook of big data technologies.* Springer International Publishing, 2017, pp. 143-178. https://doi.org/10.1007/978-3-319-49340-4_5
91. Payne N. Stirring The cauldron: Redefining computational archival science (CAS) for the big data domain. In: 2018 *IEEE International Conference on Big Data (Big Data)*, WA, USA, 10-13 December 2018, pp. 2743-2752. IEEE. https://doi.org/10.1109/BigData.2018.8622594
92. Sowe SK, Zettsu K. The architecture and design of a community-based cloud platform for curating big data. In: *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Beijing, China, 10-12 October 2013. p. 171–178. https://doi.org/10.1109/CyberC.2013.35
93. Sowe SK, Zettsu K. Curating big data made simple: Perspectives from scientific communities. *Big Data* 2014;2(1):23–33. https://doi.org/10.1089/big.2013.0046
94. McCoy MD. Geospatial big data and archaeology: Prospects and problems too great to ignore. *J Archaeol Sci* 2017;84:74–94. https://doi.org/10.1016/j.jas.2017.06.003
95. Lawrence BN, Bennett VL, Churchill J, et al. Storing and manipulating environmental big data with JASMIN. In: *2013 IEEE International Conference on Big Data*, CA, USA, 6-9 October 2013, pp. 68–75. IEEE. https://doi.org/10.1109/BigData.2013.6691556
96. Prescott A. Bibliographic records as humanities big data. In: *2013 IEEE International Conference on Big Data*, CA, USA, 6-9 October 2013, pp. 55–58. IEEE. https://doi.org/10.1109/BigData.2013.6691670
97. Affelt A. Big data, big opportunity. *Austl L Libr* 2013;21(2):78-89.
98. Chen J, Tao Y, Wang H, et al. Big data based fraud risk management at Alibaba. *The Journal of Finance and Data Science* December 2015;1(1):1–10. https://doi.org/10.1016/j.jfds.2015.03.001
99. Almalki M, Gray K, Sanchez FM. The use of self-quantification systems for personal health information: big data management activities and prospects. *Health Inf Sci Syst* 24 February 2015;3(S1). https://doi.org/10.1186/2047-2501-3-S1-S1
100. Subbiah S, Varalakshmi P, Prarthana R, et al. Energy efficient big data infrastructure management in geo-federated cloud data centers. *Procedia Comput Sci* 2015;58:151–157. https://doi.org/10.1016/j.procs.2015.08.043
101. CEOS-WGISS Data Stewardship Interest Group. Long term preservation of earth observation space data: Earth observation preserved data set content. http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf (2015, accessed 10 May 2020).

102. CEOS-WGISS Data Stewardship Interest Group. EO data preservation guidelines, v. 1.1. https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines.pdf (2015, accessed 10 May 2020).
103. Michener WK. Ecological data sharing. *Ecol Inf* 2015;29(1):33-44. https://doi.org/10.1016/j.ecoinf.2015.06.010
104. Open Geospatial Consortium. OGC standards and supporting documents. http://www.opengeospatial.org/standards (2015, accessed 9 May 2020).
105. Pouchard LC, Branstetter ML, Cook RB, et al. A linked science investigation: Enhancing climate change data discovery with semantic technologies. *Earth Sci Inf* 2013;6(3):175-185. https://doi.org/10.1007/s12145-013-0118-2
106. Goldberg D, Olivares M, Li Z, et al. Maps & GIS data libraries in the era of big data and cloud computing. *J Map Geogr Lib* 2014;10(1):100–122. https://doi.org/10.1080/15420353.2014.893944
107. Garg N, Singla S, Jangra S. Challenges and techniques for testing of big data. *Procedia Comput Sci*. 2016;85:940–948. https://doi.org/10.1016/j.procs.2016.05.285
108. Wei C, Wang XD, Ma R, et al. Analyzing on the Library Services in the age of big data. *Advanced Materials Research* 2014;1044-1045:1066–1070. https://doi.org/10.4028/www.scientific.net/AMR.1044-1045.1066
109. Yeguas V, Casado R. Big data issues in computational chemistry. In: *2014 International Conference on Future Internet of Things and Cloud*, Barcelona, Spain, 27-29 August 2014. p. 389–392. IEEE. https://doi.org/10.1109/FiCloud.2014.69
110. Zhong RY, Newman ST, Huang GQ, et al. Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Comput Ind Eng* 2016;101:572–591. https://doi.org/10.1016/j.cie.2016.07.013
111. Lichtman JW, Pfister H, Shavit N. The big data challenges of connectomics. *Nat Neurosci* 2014;17(11):1448–1454. https://doi.org/10.1038/nn.3837
112. Hey T, Tansley S, Tolle K, et al. *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research, 2009. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/
113. Lee I. Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons* 2017;60(3):293–303. https://doi.org/10.1016/j.bushor.2017.01.004
114. Fiore S, Palazzo C, D'Anca A, et al. A big data analytics framework for scientific data management. In: *2013 IEEE International Conference on Big Data*, CA, USA, 6-9 October 2013, p. 1–8. IEEE. https://doi.org/10.1109/BigData.2013.6691720
115. Jaradat M, Jarrah M, Bousselham A, et al. The internet of energy: Smart sensor networks and big data management for smart grid. *Procedia Comput Sci* 2015;56:592–597. https://doi.org/10.1016/j.procs.2015.07.250
116. Martínez-Prieto MA, Cuesta C, Arias M, et al. The Solid architecture for real-time management of big semantic data. *Future Gener. Comput. Syst.* 2015;47: 62-79. 10.1016/j.future.2014.10.016
117. Mitroff SR, Sharpe B. Using big data to solve real problems through academic and industry partnerships. *Curr Opin Behav Sci* 2017;18:91–96. https://doi.org/10.1016/j.cobeha.2017.09.013
118. Schaeffer C, Booton L, Halleck J, et al. Big data management in US hospitals: Benefits and barriers. *Health Care Manag* 2017;36(1):87–95. https://doi.org/10.1097/HCM.0000000000000139
119. Wang H, Xu Z, Fujita H, et al. Towards felicitous decision making: An overview on challenges and trends of big data. *Inf. Sci.* 2016;367–368:747–765. https://doi.org/10.1016/j.ins.2016.07.007
120. Zhou L, Pan S, Wang J, et al. Machine learning on big data: Opportunities and challenges. *Neurocomputing* 2017;237:350–361. https://doi.org/10.1016/j.neucom.2017.01.026
121. Katal A, Wazid M, Goudar RH. Big data: Issues, challenges, tools and good practices. In: *2013 Sixth International Conference on Contemporary Computing (IC3)*, Noida, India, 8-10 August 2013, pp. 404–409. IEEE. https://doi.org/10.1109/IC3.2013.6612229

122. Richards LL, Townes A, Feng YY. Curation through the Back Door: Enabling Big Data Curation Capabilities in a Non-Archival Organization. In: *Conference: Archival Education and Research Institute 2014*, PA, USA, July 2014.
123. Grady NW, Underwood M, Roy A, et al. Big data: challenges, practices and technologies: NIST Big Data Public Working Group Workshop at IEEE Big Data 2014. In: *2014 IEEE International Conference on Big Data (Big Data),* DC, USA, 27-30 October 2014, pp. 11–15. IEEE. https://doi.org/10.1109/BigData.2014.7004470
124. Edwards JS, Taborda ER. Using knowledge management to give context to analytics and big data and reduce strategic risk. *Procedia Comput Sci* 2016;99:36–49. https://doi.org/10.1016/j.procs.2016.09.099
125. Fister B. Big data or big brother? Data, ethics, and academic libraries. *Library Issues: Briefings for Faculty and Administrators* March 2015;35(4).
126. Gorton I, Xu R, Yang Y, et al. Experiments in curation: towards machine-assisted construction of software architecture knowledge bases. In*: 2017 IEEE International Conference on Software Architecture (ICSA),* Gothenburg, Sweden, 3-7 April 2017, pp. 79–88. IEEE. https://doi.org/10.1109/ICSA.2017.27
127. Agrawal D, Abbadi AE, Arora V, et al. Mind your Ps and Vs: A perspective on the challenges of big data management and privacy concerns. In: *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, Jeju, South Korea, 2 April 2015, pp. 1-6. https://doi.org/10.1109/35021BIGCOMP.2015.7072814
128. Budhiraja R, Thomas R, Kim M, et al. The role of big data in the management of sleep-disordered breathing. *Sleep Med. Clin.* 1 June 2016;11(2):241–255. https://doi.org/10.1016/j.jsmc.2016.01.009
129. Liu Y, Qiu M, Liu C, et al. Big data challenges in ocean observation: A Survey. *Pers. Ubiquitous Comput.* 2017;21(1):55–65. https://doi.org/10.1007/s00779-016-0980-2
130. Howe III EG, Elenberg F. Ethical challenges posed by big data. *Innov Clin Neurosci* 2020;17(10-12):24-30. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7819582/
131. Elliot M, Mackey E, O'Hara K, et al. The anonymisation decision-making framework. *UKAN Publications*. http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf (2016, accessed 9 May 2020).
132. Larson E. Big questions: Digital preservation of big data in government. *The Am. Arch.* 2020;83(1):5-20. https://doi.org/10.17723/0360-9081-83.1.5
133. Knight M. Data Curation 101: The what, why, and how. *Dataversity*. https://www.dataversity.net/data-curation-101/ (2017, accessed 9 May 2020).
134. Tang Y, Pichler K, Füllgrabe A, et al. Ten quick tips for biocuration. *PLoS Comput Biol* 2019;15(5);e1006906. https://doi.org/10.1371/journal.pcbi.1006906
135. Lowndes JSS, Best BD, Scarborough C, et al. Our path to better science in less time using open data science tools. *Nat Ecol Evol* 2017;1. https://doi.org/10.1038/s41559-017-0160
136. Dietrich D, Adamus T, Miner A, et al. De-mystifying the data management requirements of research funders. *Issues Sci. Technol. Librariansh*. 2012:70. https://doi.org/10.29173/istl1556
137. Garoufallou E, Gaitanou P. Big data: opportunities and challenges in libraries, a systematic literature review. *Coll res libr* 2021;82(3). https://doi.org/10.5860/crl.82.3.410
138. Hoy MB. Big data: An introduction for librarians. *Med. Ref. Serv. Q.* 2014;33(3):320–326. https://doi.org/10.1080/02763869.2014.925709
139. Zhan M, Widen G. Public libraries: Roles in big data. *The Electronic Library* 2018;36(1):133-145. https://doi.org/10.1108/EL-06-2016-0134