IJDC | *Peer-Reviewed Paper*

Revisiting the Data Lifecycle with Big Data Curation

Line Pouchard
Purdue University Libraries

Abstract

As science becomes more data-intensive and collaborative, researchers increasingly use larger and more complex data to answer research questions. The capacity of storage infrastructure, the increased sophistication and deployment of sensors, the ubiquitous availability of computer clusters, the development of new analysis techniques, and larger collaborations allow researchers to address grand societal challenges in a way that is unprecedented. In parallel, research data repositories have been built to host research data in response to the requirements of sponsors that research data be publicly available. Libraries are re-inventing themselves to respond to a growing demand to manage, store, curate and preserve the data produced in the course of publicly funded research. As librarians and data managers are developing the tools and knowledge they need to meet these new expectations, they inevitably encounter conversations around Big Data. This paper explores definitions of Big Data that have coalesced in the last decade around four commonly mentioned characteristics: volume, variety, velocity, and veracity. We highlight the issues associated with each characteristic, particularly their impact on data management and curation. We use the methodological framework of the data life cycle model, assessing two models developed in the context of Big Data projects and find them lacking. We propose a Big Data life cycle model that includes activities focused on Big Data and more closely integrates curation with the research life cycle. These activities include planning, acquiring, preparing, analyzing, preserving, and discovering, with describing the data and assuring quality being an integral part of each activity. We discuss the relationship between institutional data curation repositories and new long-term data resources associated with high performance computing centers, and reproducibility in computational science. We apply this model by mapping the four characteristics of Big Data outlined above to each of the activities in the model. This mapping produces a set of questions that practitioners should be asking in a Big Data project.

Received 29 January 2015 | *Revision received* 16 July 2015 | *Accepted* 22 April 2016

Correspondence should be addressed to Line Pouchard, Purdue University Libraries, 504 W State Street, West Lafayette, IN 47906, USA. Email: pouchard@purdue.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v10i2.0

Introduction

As science becomes more data-intensive and collaborative, researchers increasingly use larger and more complex data to answer research questions. The capacity of storage infrastructure, the increased sophistication and deployment of sensors, the ubiquitous availability of computer clusters, the development of new analysis techniques, and larger collaborations allow researchers to address "grand challenges"[1] in a way that is unprecedented. Examples of such challenges include the impact of climate change on regional agriculture and food supplies, the need for reliable and sustainable sources of energy, and the development of innovative methods for the treatment and prevention of infectious diseases. Multi-disciplinary, sometimes international teams meet these challenges and countless others by collecting, generating, cross-referencing, analysing, and exchanging datasets in order to produce technologies and solutions to the problems. These advancements in science are enabled by Big Data, recently defined as a cultural, technological, and scholarly phenomenon (Boyd and Crawford, 2012).

The coinage 'Big Data' has multiple etymological and picturesque origins, discussed by Lohr (2013) and the blog posts he elicited. A 1989, the non-fiction journalist and author, Erik Larson, gives a portentous definition that is not computer related in an article about junk mail:

> "The keepers of Big Data say they do it for the consumer's benefit. But data have a way of being used for purposes other than originally intended" (Larson, 1989).

In the mid-1990s, the phrase appears to have been used a lot around Silicon Graphics, both in academic presentations and sales pitches to scientists, customers, analysts and press[2]. Around this time, an early academic definition appears in a paper found in the ACM Digital Library, as data that is too large to fit into local computer memory and is tied to the demands of computational fluid dynamics and visualization: "visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk: we call this the problem of Big Data." (Cox and Ellsworth, 1997). These early definitions encompass characteristics of Big Data that have remained important to this day: the mutual interaction between computer advances and the demands of science, and the plasticity of data.

The three Vs that characterize Big Data – volume, velocity, and variety – were first proposed by Doug Laney, now a Gartner consultant in the context of emerging business conditions (2001). They were later expanded to include veracity. In the Laney report, volume refers to the increased amount of information that an enterprise can accumulate about a transaction using e-commerce channels, up to ten times more than previously, thus requiring increased storage and strategies for selective data collection. Velocity is the speed at which this information accumulates, where the appropriate harnessing of speedy information confers a competitive advantage. Beyond physical bandwidth and protocols, enterprises deploy pointed solutions that balance the requirements of decision

1 21st Century Grand Challenges – Office of Science and technology Policy: http://www.whitehouse.gov/administration/eop/ostp/grand-challenges

2 "See: http://en.wikipedia.org/wiki/Talk:Big_data

cycles and data latency without assuming the entire data chain must be real-time. Variety points to "incompatible data formats, non-aligned data structures, and incompatible semantics" as the greatest barriers to be overcome. While later technological developments in handling volume have progressed in great strides, the challenges of variety still require a great deal of attention.

In parallel, research data repositories (RDR) have been built to host research data in response to the requirements of sponsors that research data be publicly available. However, the RDR landscape is heterogeneous in terms of infrastructure, strategies for permanent access, re-use of data, and funding models (Pampel, Vierkant, Scholze et al., 2013). Data management and curation services are becoming more common among research libraries that are re-inventing themselves to respond to a growing demand to manage, store, curate and preserve the data produced in the course of publicly-funded research (Soehner, Steeves and Ward, 2010). As research data managers and librarians are developing the tools and knowledge they need to meet these expectations, they inevitably encounter the conversations around Big Data. Big Data is poised to affect every layer of society, as companies and governments now have the means to collect huge amounts of data on every person and every action in every walk of life, and methods for analyzing this data are more effective. In research, a potential paradigm shift is taking place in both the humanities and social sciences, as well as science and engineering disciplines, as issues such as collection gaps, metadata, interpretation, and use raised by the new rubric of Big Data require new theoretical underpinnings (Boelstorff, 2013).

This paper explores some definitions of Big Data and discusses their associated issues, taking the perspective of their impact on data management and curation. To explore the issues encountered by librarians in the conversations about Big Data, we take the framework of the research data life cycle model. This allows us to perform two tasks. The first is to examine and compare two data life cycle models that have been produced within the context of Big Data projects. The second is to take the various characteristics of Big Data and examine the issues raised at each step of the life cycle. This will allows us to propose a new data life cycle that fits the characteristics of Big Data and map each activity to these characteristics. Questions to explore at each step are also provided.

Definitions: The Characteristics of Big Data

Big Data gained momentum as a phenomenon in scientific research with a series of white papers from the Computing Community Consortium starting in 2008[3]. Bryant, Katz and Lazowska (2008) bring together examples from science, commerce, medicine, and national security that illustrate the extent to which large, complex datasets are accumulated thanks to new technologies that include sensors, distributed computer systems, data storage and high performance networks.

Also in 2008, various contributors to the journal *Nature* were asked to speculate about the technologies and trends most likely to have an impact on society in the next ten years in a special issue titled 'Big Data: Science in the Petabyte Era'. In it Clifford Lynch examines the costs and challenges of data stewardship in the long term, noting that funding agencies and educational institutions are equally reluctant to take on this

3 Computing Community Consortium white papers: http://www.cra.org/ccc/visioning/ccc-led-white-papers

responsibility. According to Lynch, these challenges will ultimately be met by university consortia and focused archives, with the best stewardship coming from an engagement between preservation institutions and disciplinary focus. In the meantime, general purpose data management as provided by libraries will have an important role but it will also have its limits (Lynch, 2008).

Volume or size of the data has been a thorny issue from the start. The question of 'how big is big?' seems to be a moving target. Consensus seems to emerge around the fact that volume/size characterizes Big Data as that which exceeds the capacity of what can be stored with conventional means, and what seems big today will be small tomorrow (Ward and Barker, 2013). Conventional means of storage include databases that store data into relational tables. Non-conventional storage architectures include the No SQL systems, such as Apache Hadoop, that provide a different mode of organization, e.g. document stores, graph-based storage, or key-value stores. The architecture of these systems enables faster transaction and retrieval rates, as well as the ability to expand storage by simply adding new storage nodes.

Variety or complexity features prominently among the characteristics of Big Data. Big Data often refers to the vast amounts of unstructured data, such as tweets, videos or images, such as medical images. In 2012, it was estimated that 85% of all data is unstructured and generated by humans (Mills et al., 2012). The variety of formats and sources underlies the complexity of the analysis, as data from numerous, heterogeneous sources must be processed and adequately integrated prior to analysis, especially in commercial enterprises. In biology, data is produced by a large variety of experiments that produce genetic sequences, interaction of proteins or findings in medical records: this data is much more heterogeneous than in physical sciences (Marx, 2013).

Velocity, the speed at which data accumulates, including its rate of change, presents challenges for storage, access, and analysis. Speed creates flows of data that need to be managed, organized and analyzed within timeframes that re-define real-time. Depending on the enterprise and the purpose for which data is collected, decisions based on the data may have to be made within a 24 hour time-frame or milliseconds, as in the stock market. This may apply to airline prices, election result projections, or the discovery of new celestial objects in one part of the world that need to be confirmed in another before the end of the night. With Big Data, it is more advantageous to move compute power and processing algorithms to the data than bring data to the computing nodes.

The three Vs of Big Data were expanded by IBM and others with the concept of veracity[4]. Veracity refers to the quality of Big Data, understood in terms of accuracy (reliable methods of data acquisition), completeness of data (are there duplicates or missing data?), consistency (are measurements and unit conversions accurate?), uncertainty about its sources, and model approximations (Lukoianova and Rubin, 2014). Big Data can also be full of errors, or noise, such that its analysis can become meaningless. For instance, Big Data may be prone to overfitting, a case when the learning algorithms used to analyze existing data are not robust to noise and lead to inaccurate predictions. In order for Big Data to yield the insights it is expected to, users must be able to trust the data and its transformations. Data-driven decisions emphasize the need for traceability and provenance.

4 Infographic: The four Vs of Big Data: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Data Life Cycle Models

Data life cycle models present a structure for organizing the tasks and activities related to the management of data within a project or an organization. They also present a means of communicating these tasks to the intended audience, which may include researchers, data managers, curators, repository specialists, librarians, and project managers in charge of organizing data in a project or in a lab. This diverse audience results in a wide variety of data life cycles, with some focused on an organization, others on individual researchers, and others on the community at large. The interest of these models lies in the tension between the ideal representation of the model and its practical applications. The Committee on Earth Observation Satellites Working Group on Information Systems and Services documents 52 data life cycles (CEOS, 2012). While it is out of the scope of this paper to analyze these, we can present the results of two recent studies of various data life cycle models. We then supplement these analyses with an examination of the scientific data life cycle from DataONE[5] (Figure 1) and the Data Life Cycle Laboratories (DLCLs) (Figure 2) (van Wezel et al., 2012).

Ball (2010, 2012) provides an analysis of eight data life cycles, helpful within the context of organizations (Australian National Data Service, UK Data Archive), research projects (I2S2, Research360 Institutional Research Data lifecycle, Capability Maturity Model for Scientific Data) or within a specific community (DCC, DataONE, DDI Combined Life Cycle Model). These help map out the tasks and issues encountered in the process of data management and curation. Some of these life cycles are presented from the perspective of a researcher engaged in scientific research through a project, often supported by a grant. Others are presented from the point of view of an organization or data center specializing in data curation and preservation. Some are designed for a specific community, such as data managers planning their tasks. Ball finds that the life cycles present a simple view of activities that generally include data generation, collection and a processing stage that may include various forms of transformation, analysis, and dissemination. He points out that this view obscures the complexity and variety of the research process, and does not represent the early stages of this process with the same amount of details. We will use these observations to compare various models.

Carlson (2014) classifies seven life cycle models for their use in determining the data services that libraries may offer. He distinguishes between models representing data management (where data is represented during the active research phase) and data curation (where data is shared to a larger group of users than its original creators and is frequently under the stewardship of a third party). We will also use this distinction to compare various models. Life cycles are used as a means of communication to their intended audience. By mapping services to various steps of a data life cycle, gaps in services are identified, and informed decisions about which services to offer and how to scope these services can be made. Carlson distinguishes between several types of data life cycle models, such as individual-based, community-based, and organizational. He notices that models tend to represent data-related activities in an orderly and linear fashion, which is rarely the case in reality. Second, models tend to overlook the diversity of approaches and practices that may be present, a point also made by Ball. Finally, models tend to reflect the biases of the organizations that created them, and may not be easily adaptable.

5 DataONE life cycle model: https://dataone.org/data-life-cycle

 doi:10.2218/ijdc.v10i2.0

The Curation Lifecycle Model produced by the Digital Curation Centre (DCC) is an example of a community-based model (Higgins, 2008). The DCC focuses on data curation. Its purpose is to address the needs of the community when organizing the steps identified in data curation and preservation. The fourth level of the DCC model describes the various stages of the life cycle that should have a curation component associated with them. These include: conceptualize, create or receive, appraise and select, ingest, preservation action (including quality control), storage, access, use and re-use, and transform (migrate the data in case of technology obsolescence).

We now turn to two lifecycle models conceived within the context of Big Data. DLCLs are five community-specific initiatives of the Large Scale Data Management and Analysis (LSDMA)[6] project of the Helmholtz Association of research centers in Germany. Each initiative provides domain-specific data analysis tools and cross-cutting data management services, and optimizes the scientific data life cycle for the community it serves (van Wetzel, et al., 2012, and Jung et al., 2014). These research communities include: energy (smart grids, battery research, and fusion research), earth and environment (climate model and earth observation satellite data), health (the virtual human brain map), key technologies (synchrotron radiation, nanoscopy, high throughput microscopes, electron-microscope imaging techniques), and structure of the matter (large instruments, heavy ion research, elementary particle physics).

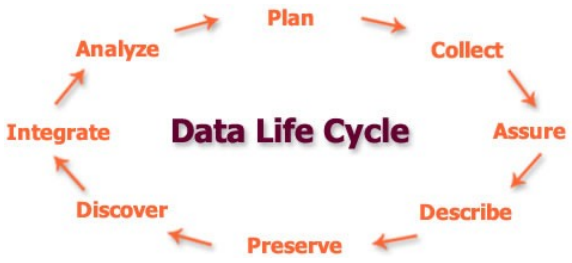

Figure 1. DataONE Data Life Cycle model.

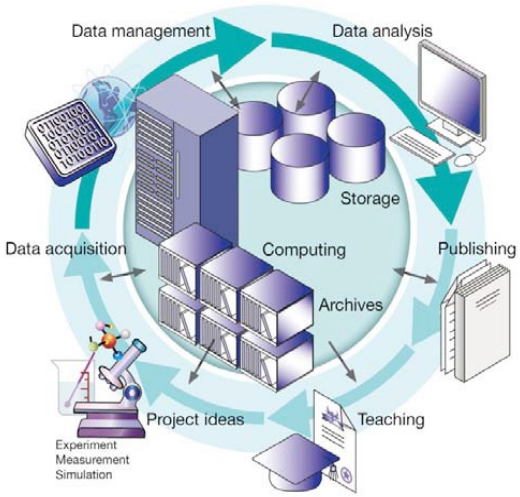

Figure 2. Data Life Cycle Labs (DLCL) model.

6 LSDMA: http://www.helmholtz-lsdma.de

The model used by the DLCLs starts the cycle at the bottom with project ideas. Data generated by measurements, observations, experiments and simulations are used to generate hypotheses. As a result, raw data needs to be preserved. Hypotheses lead to planning for new research sometimes using design of experiments. The data acquisition phase is conducted by teams of scientists operating large specialized instruments, e.g. telescopes, neutron sources, or running simulations on high performance computing systems, e.g. global climate simulations coupling individual constituent simulation models, or aggregating large collections of heterogeneous data. Data management tasks and the procurement and maintenance of infrastructure in common between DLCLs are conducted by the Data Service and Integration team. These tasks include distributed data management, storage and access, metadata and ontologies for data identification and derivation over time, standardization of formats, data security and high performance analysis (van Wetzel, 2012; Jung et al., 2014).

The DataONE Data Life Cycle model is a community-based model designed to organize the best practices for curating and preserving the Earth Science data found in DataONE. Supported by the National Science Foundation, DataONE is a federation of data centers and archives comprised of a) a cyber-infrastructure enabling distributed access to Earth Science observations through metadata discovery, and b) a community of practitioners, including sponsors, researchers, students, librarians, and citizen scientists. DataONE ensures the curation and preservation of Earth Science observations for long-term use, as well as educating its stakeholders in best practices for data management and curation. The model combines different workflows with some fluidity in the order of occurrence (Ball, 2010). If data is to be generated by instruments and sensors, the DataONE steps of an investigation include: plan, collect, assure, describe, preserve and analyze. In the case of re-use of existing data, the steps involve discover, integrate, assure and analyze.

We present a table indicating various aspects of these life cycle models to illustrate gaps in the presented activities. We use the DCC, DLCL, and DataONE models for this comparison. We use the analysis found in the literature and discussed above to provide axes of comparison. The models were chosen because the DLCL and the DataONE models have been developed in the context of Big Data in the following way: DLCLs present volume and complexity of the data in each of its scientific initiatives. DataONE illustrates variety in the Earth Sciences as it gathers data from many heterogeneous sources, each a data center with its own policies, data sources, and search and analysis tools (Hampton et al., 2013). We also use the DCC model as a reference point. As we can see, no model presents all categories.

Table 1. Comparison of data life cycles.

	DCC	DLCL	DataONE
Integrates the research process into the data life cycle model (data management)	No	Yes	Yes
Focuses on data curation and preservation activities (data curation)	Yes	No	No
Models the early stages of the cycle, including data preparation in details	No	No	No

The DLCL model has been developed in the context of strategically funded research centers that pool resources to provide tools to scientists in the Helmholz Association. While the DLCL captures the research process it does not take the perspective of data curation. Only 'archives' is mentioned in the bottom right of the cycle and it is not clear if this label encompasses the infrastructure or the practices involved in archiving. In addition, several activities seem to be lumped under the label 'Data Management' without adequate description or differentiation. Other activities that may fall under data management are singled out. The DLCL and DataONE models take a different but complementary perspective on the scientific data life cycle. The DLCL model focuses on a researcher's view supported by infrastructure but does not address data curation.

Although it was developed in the context of Big Data, the DataONE Data Life Cycle Model does not account for Big Data in several ways. First, the preparation of data for analysis, a crucial step in working with Big Data, is missing from this life cycle. Data undergoes numerous transformations as they are prepared for analysis and checked for quality. This phase, sometimes called "data wrangling" when working with Big Data and visualization, is often reported as very tedious and time-consuming (Kandel, Heer et al., 2011). Second, in this life cycle, the 'Assure' and 'Describe' steps occur sequentially at specific phases of the model. As described later in this paper, Assure and Describe tend to occur as ongoing activities to ensure quality and metadata collection. Finally, as our discussion below shows, the 'Preserve' activity tends to occur following the analysis, as this activity is tied to the reproducibility of research.

This analysis highlights gaps in the data life cycle models in areas that are important for Big Data. It leads us to propose a new data life cycle for use with Big Data projects.

The Big Data Life Cycle Model

We propose a Big Data Life Cycle Model (Figure 3) that combines the perspective of research with that of data curation, identifies the tasks of data management that lead to analysis, while preserving the curation aspect, and encompasses the steps necessary to handle Big Data. We first discuss each activity, then we apply this model by providing mappings between each activity and the Big Data characteristics outlined above (the four Vs of Big Data). On the surface, Big Data curation has many similarities with other types of curation. But Big Data raises many issues for each step of the management and curation process that are not adequately described in existing models. While the labels may be similar, the activities encounter different challenges.

The Big Data Life Cycle is intended as a general model to be used by researchers, data managers, and librarians to plan the workflow of research data management and data curation activities in their projects or organizations. It is not intended to replace a model serving a specific community with programmatic funding and associated compliance mechanisms, like the DLCL labs. Practitioners can use the Big Data Life Cycle Model to determine the phase their project or their data currently is in, and the questions to be asked at that phase. It can be used in an iterative manner with a project entering the life cycle at any point. Datasets in the same project may be situated at different phases of the cycle. They would thus be submitted to different tasks. In addition, tasks in a project do not necessarily complete the phases in a cascading order nor do all tasks necessarily apply to all data. It is rarely the case when a phase occurs only once in a project. For instance, data may be acquired at several points in the duration of a project. Questions related to acquisition will apply to the datasets acquired

at that point. Other datasets may already be in the analysis phase. The questions related to analysis will then be applicable to those. These questions are addressed in the next section.

We now explain every activity presented in the figure of the Big Data Life Cycle (Figure 3). The 'Describe' and 'Assure' activities are presented outside the cycle to emphasize that they should be present at every step of the Big Data Life Cycle. The infrastructure supporting the model is represented by cogs at the center of the figure and is also discussed.

The 'Describe' Activity: Describing the data and processes used in the analysis at every step – capturing the provenance trace – is crucial for Big Data. The earlier curation-related tasks are being planned in the data management life cycle, the easier they may be to execute (Ball, 2010). Take for instance the collection of metadata for datasets. The earlier in the research cycle researchers start collecting metadata, the closer metadata is to data acquisition and the more likely it reflects the source and facilitate discovery. In addition, it is less likely that a potentially crucial transformation of the data will be omitted if it is documented. Designing a naming scheme for potentially millions of files is part of these tasks. The naming scheme may include a timestamp and keep track of different versions of the same data.

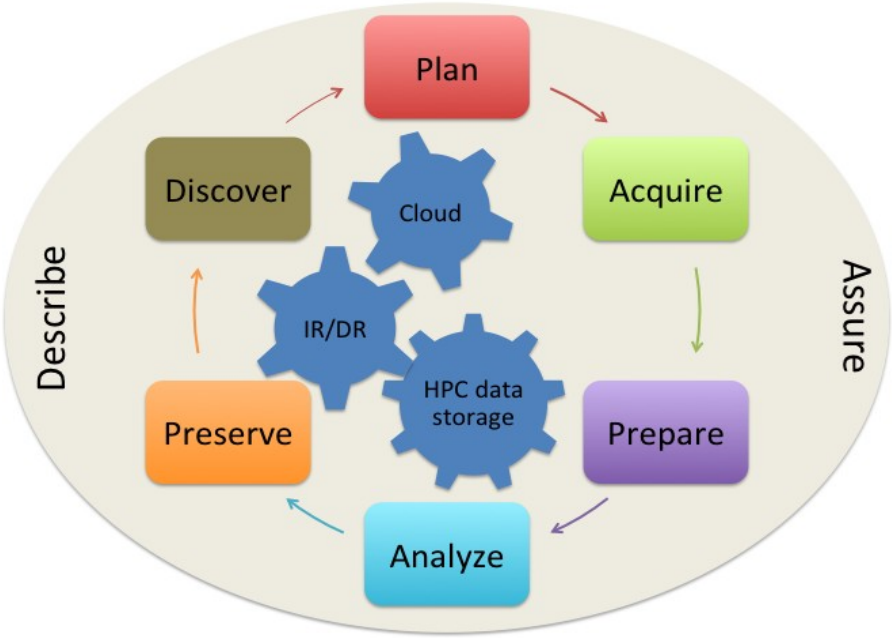

Figure 3. The Big Data Life Cycle Model. The background labelled 'Assure' and 'Describe' highlights that these activities take place at every step. The central cogs represent the infrastructure supporting data held in cloud infrastructure, an institutional repository (IR), a disciplinary repository (DR), or a high performance computing center (HPC) data facility.

Documenting data sources, experimental conditions, instruments and sensors, simulation scripts, processing of datasets, analysis parameters, thresholds, and analysis methods ensures not only much needed transparency of the research, but also data discovery and future use in science. Documenting variables, transformation processes, workflows and analysis decisions are also important. In industry, this documentation

 doi:10.2218/ijdc.v10i2.0

provides a basis and a justification for decision-making. The amount and granularity of metadata generated by Big Data explorations require more than metadata standards: scalable tools for the automatic generation and extraction of metadata are needed so that the relationships between raw data and analysis results are preserved. Semantic tools that derive metadata from annotations and ontologies can help. Automated scientific workflow tools that capture the steps needed to obtain results also help. Some open source tools exist and are scalable[7]. They require tight integration into computing workflows to capture provenance and workflow history and are suited to some scientific inquiries.[8]

The "Assure" activity: Describing the data, including its processing and analysis, from the start of the project plays a crucial role toward assuring the quality of the data. However, it's not in itself sufficient. Assuring the quality of data includes quality assurance, which may occur once a dataset is analyzed, and quality control, a pro-active process with procedures in place to ensure quality. Data quality is directly related to the veracity characteristic of Big Data, as accuracy, completeness, and uncertainty about sources play a crucial role in veracity. Similar to data documentation, issues of quality arise at every stage of the life cycle, including acquisition, preparation, analysis, and preservation (Sukumar, Natarajan and Ferrell, 2014). Flagging missing data, documenting unit or format conversions, entity resolution, and documenting a data source go a long way toward assuring trust in the data. Acquiring data from multiple sources presents its own quality-related challenges. Data from various sources may be outdated or conflicting. Sources may have different levels of quality resulting in a combined dataset with the lowest common denominator. Different models of data representation may lead to metadata errors and erroneous conversions, such as with dates and geographical locations. In addition, errors tend to propagate as the number of data integration steps and processes increases.

The 'Planning' activity: The selection of data for preservation must be discussed at the planning stage due to the potential volume of data to be preserved. Keeping all raw data may be required, as some experiments are too costly to reproduce (e.g. examining the properties of new materials using a neutron beam), whilst others are bed on transient observations. In some cases the volume and velocity of data preclude the preservation of raw data. In these cases, data is analysed "on the fly" and only analysis results are preserved, thus preventing forensic analysis: monitoring for failure with voltages, temperatures, and power consumption on each core, memory bank, and network chip of a supercomputer with a quarter million cores is such an example. In other cases, it is cheaper and easier to run a simulation or a sequencer again to obtain the raw data than to preserve it.

The "Acquire" activity: The 'acquire' activity reflects how data is produced, generated, and ingested in the research process. Data acquisition may be the result of using remote sensors, instruments such as mass spectrometers and sequencers, or it may be the result of computational simulations or downloads from external sources such as a disciplinary repository or the Twitter API (Application Programmer's Interface). In the case of sensor data, much of the raw data can be compressed and filtered so that only useful data is selected for preservation; the challenge is in designing these filters. (Labridinis and Jagadish, 2012).

The 'Prepare' activity: Preparing datasets and staging them for analysis is a time-consuming step with Big Data and its complexity is often overlooked. Data wrangling may involve reformatting, cleaning, and integrating data sets so that they are amenable

7 Tools include Kepler: https://kepler-project.org

8 See https://kepler-project.org/users/projects-using-kepler

to analysis and visualization (Heer and Shneiderman, 2012). It includes designing customized scripts written in Python or R to normalize the datasets, reconcile date formats and geographical coordinate systems, remove duplicates, split up columns, supply headers, and generally make the dataset usable for the analysis program. Kandel and Heer (2011) define data wrangling as a process of iterative data exploration and transformation that enables analysis. If datasets from multiple sources are required for analysis, data from various sources must be integrated and pipelines of data processing must be built, with the data output of one or several processes becoming the input of another. This is particularly the case in bioinformatics where sequences, annotations, pathways, transcription factors from numerous data sources are used to better understand diseases. In other Big Data explorations, such as social media analysis for marketing, preparing the data may include integrating text data with geo-referencing data. Processing the data may require integrating text from tweets and blogs, stemming and normalizing data using natural language processors, and annotating them with ontology entities.

The 'Analysis' activity: The 'analysis' activity is the domain of the scientists performing research. Statistical methods and machine learning, in particular, feature prominently with Big Data. However, recording and preserving the parameters of experiments, including simulation scripts, and the entire computational environment are needed for the reproducibility of results (Stodden, 2010). Reproducibility is the ability to repeat an experiment to the degree necessary to assess the validity and importance of the claimed results (James, 2014). Objects that have not traditionally been part of data curation, such as software and source code, may need to be selected for preservation. Although popular software repositories, such as GitHub, offer the possibility of assigning a Digital Object Identifier (DOI) to source code, the process involves archiving your source code into Zenodo, a science repository hosted on the cloud infrastructure of CERN's Large Hadron Collider and funded by the European Commission. One advantage of preserving code is the relatively small volume compared to data. With Big Data, hardware choices, software updates, and configuration changes are typically not controlled by the researcher but by the facility providing computational power. Additional information is needed, such as a description of its features, the version of underlying libraries or toolkits used in producing the code, and the range of meaningful values for input parameters. Source code and the underlying architectures age very fast, thus presenting new challenges to curation.

The 'Preserve' activity: In order to preserve results for long-term use, data life cycle models should not just capture activities performed when researchers are ready to publish their data. The preservation activity should include the creation of pipelines or workflows that track dependencies between data and processes, and allow linking raw data to results in a publication. Preservation activities should aim to capture data transformations in order to address the challenges of Big Data, possibly along the lines of recording processes into formal process management plans as described in Miksa, Strodl and Rauber (2014). Preserving Big Data for the long-term is about preserving many series of processes that are interconnected and may be repeated several times during the research lifecycle. Preserving processes is more difficult to achieve because processes and inputs/outputs change over the course of time, possibly within days or even hours, in the case of Big Data. In this context, it is more accurate to talk about data sets or data objects, rather than data, as they refer to discrete, quantifiable entities that are transformed by processes.

The "Discover" activity: The 'discover' activity refers to the set of procedures that ensures that datasets relevant to a particular analysis or collection can be found by other

than those involved in the project. At this stage, a researcher must decide which data will be made discoverable. Discovery is made possible by data sharing, a practice that has not yet gained wide-spread acceptance in all disciplines (Tenopir et al., 2011). Data sharing is required by funding agencies, such as the National Science Foundation and the National Institutes of Health in their public access plans. Disciplinary and community-based repositories, one of the central cogs in the diagram, provide important infrastructure for Big Data discovery. These repositories rely upon data and metadata standards, federated searches, software tools and persistent identifiers that facilitate data discovery (Michener, 2015). Semantic searches supported by ontologies that provide entities for query expansion and metadata annotations also enhance discoverability for Big Data (Pouchard et al., 2013). Ontologies allow structuring information into networks of classes that can be searched using relationships such as proximity, synonymy, location, and others.

The computing infrastructure that supports Big Data (the cogs in the diagram) has been an enabler in the explosion of Big Data and for the types of research questions that can be asked with it. Most often, institutional repositories that provide data curation services at a research institution are not well equipped to handle the volume and variety of data that occur in some Big Data projects. Mechanisms for ingesting multiple formats and converting them to formats appropriate for archiving are in place. Metadata, description of datasets and checksums to detect data corruption are common procedures that reinforce the quality of the data. However, volume, including size of individual datasets and total amount of datasets, may become a challenge beyond the capacity of institutional repositories. Data transfer requiring high bandwidth may also become a bottleneck. To accommodate those needs, high performance computing centers on campuses are starting to offer high capacity storage and fast access resources to accommodate Big Data beyond the life of a project. These resources are not typically equipped for data management and curation. As a result, strategic partnerships between libraries and HPC centers are emerging, where one partner provides the infrastructure and the other the data curation expertise.[9] As the preservation of the computing environment is one of the best practices for reproducibility of research, HPC centers need to take more responsibility in facilitating these practices (Fahey and McLay, 2014). Partnering with data curation specialists is one way to achieve this goal. Furthermore, institutional repositories and these new resources have to define their respective place in the continuum of Big Data research. One possible direction is to focus on datasets backing up a publication for institutional repositories, and project datasets for the HPC data resources. This would include defining workflows for migration from one resource to the other at the time of publication.

Application of the Big Data Life Cycle

In order to illustrate the use of the Big Data Life Cycle model, we apply it to the four characteristics of Big Data defined at the beginning of the paper (the four Vs). By mapping these characteristics to each step of the life cycle in Table 2, we obtain questions and issues that a data manager or library practitioner can ask when faced with a Big Data project. The answers to these questions will help make decisions for the infrastructure as well as guide the activities of data management of a Big Data project.

9 For examples of such strategic partnerships, see http://www.sdsc.edu/News%20Items/PR040912_chronopolis.html and https://www.rcac.purdue.edu/services/data/

Table 2. Issues raised by the characteristics of Big Data applied to the Big Data life cycle.

	Volume	Variety	Velocity	Veracity
Plan	What is an estimate of data volume and growth rate?	How do data policies from different sources combine? What provisions are made to accommodate sensitive data?	Are bandwidth and planned storage sufficient to accommodate input rates?	What are the data sources? What allows a researcher to trust them? Who will own derived data, and data resulting from aggregation?
Acquire	What is the most suited form of storage (databases, NoSQL, cloud)?	What are the data formats? What steps are needed to integrate data from different sources?	Will datasets be aggregated into series? Will metadata apply to individual datasets, to series, or both?	Who collects the data? Do they have the tools and skills to choose the best available sources?
Prepare	What are the implications of volume in the preparation of datasets for analysis?	Are different types of workflows needed to process data from different sources? Do we need to remove blanks, duplicates, split columns, add/remove headers?	Are different schemes for file naming required? Do some data need to be discarded due to accumulation?	Are the wrangling steps sufficiently documented to foster trust in the analysis?
Analyze	Are adequate computing power and analysis methods available?	Are the various analytical methods compatible with the different datasets?	At what time point does the analytical feedback need to inform decisions?	What level of code sharing is needed to ensure transparency and reproducibility? Is the chosen type of analysis appropriate for the collected data?

Table 2. Issues raised (continued)

 doi:10.2218/ijdc.v10i2.0

	Volume	Variety	Velocity	Veracity
Preserve	Should raw data be preserved? What storage space is needed in the long-term? Who will provide it?	Are there different legal or policy considerations regarding sharing for each data source? Are there conflicts with privacy and confidentiality?	When does data become obsolete?	What are the trade-offs if only derived products and no raw data are preserved?
Discover	What part of the data (derived, raw, software code) will be made accessible to searches? How will a potentially large number of results be displayed?	What search methods best suit this type of data? Keyword-based, geo-spatial searches or metadata-based, semantic searches?	What degree of search latency is tolerable?	Providing well documented data in open access allows transparency. How is veracity supported with sensitive data that cannot be shared without restrictions?

The cells within the tables highlight which issues are of importance at each stage, based on the characteristics of Big Data. The answers to these questions or resolutions of the issues will depend on individual projects, but the questions remain applicable across the board. Not every Big Data project will encounter all the questions, and some projects may exhibit some features more prominently than others. The issues highlighted by the four Vs in Table 2 make it clear that, while focused on curation and preservation as the long-term goal, a data life cycle for Big Data must start with the planning and conceptualizing of a project: more details must be recorded due to data complexity, heterogeneity, and sheer number of processing steps. Volume and velocity are potential bottlenecks for the infrastructure that must be planned from the start. Deciding which datasets (raw, cleaned-up, derived) are to be preserved must also be decided as early as possible in the data life cycle as it impacts the infrastructure.

Conclusion and Future Work

In this paper we proposed a new life cycle model that specifies the various phases of the research process when dealing with Big Data. We structured the Big Data Life Cycle around the activities of planning, acquiring, preparing, analyzing, preserving and discovering the data, with describing data and assuring quality as pervasive activities throughout all phases. We investigated two previous data life cycle models created in the context of Big Data projects and found them wanting. We mapped the four Vs of Big Data (volume, variety, velocity, and veracity) to each phase of the life cycle to highlight the issues raised for data management and curation by each characteristic. These mappings provided a set of questions that a data practitioner may want to ask at

each step of a Big Data project. The mappings also allowed us to confirm that data curation for Big Data should start at the beginning of the project, as questions related to data selection, data preparation, metadata and workflows impact curation decisions.

We are working on applying this Big Data Life Cycle model to various Big Data projects at a land grant university. By applying the model to each of these projects and asking the questions related to each Big Data characteristic, we will be able to elicit the specificity of each project. In addition, we will be able to evaluate the model's usefulness in practice, as well as measure researchers' engagement with data management in recent Big Data projects. Looking toward the future, the phrase Big Data may fade away like other catchall phrases before it, but its characteristics and their implications for data curation will not.

References

Ball, A. (2010). Review of the state of the art of the digital curation of research data. Bath: University of Bath. Retrieved from Opus: University of Bath Online Publication Store website: http://opus.bath.ac.uk/19022/

Ball, A. (2012). Review of data management lifecycle models. Retrieved from Opus: University of Bath Online Publication Store website: http://opus.bath.ac.uk/28587/

Boelstorff, T. (2013). Making Big Data, in theory. *First Monday, 18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society, 15*(5), 662-679. doi:10.1080/1369118X.2012.678878

Bryant, R., Katz, R.H., & Lazowska, E.D. (2008). Big-data computing: Creating revolutionary breakthroughs in commerce, science and society: December. Retrieved from http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf

Carlson, J. (2014). The use of life cycle models in developing and supporting data services. In *Research Data Management: Practical Strategies for Information Professionals* (pp. 63). Purdue University Press.

Committee on Earth Observation Satellites (CEOS) Working group on Information Systems and Services (2012). Data life cycle models and concepts, CEOS Version 1.2. Retrieved from http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS_DSIG-Data-Lifecycle-Models-and-Concepts-v8_Sep2011.docx

Cox, M. & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings, Visualization '97*. doi:10.1109/VISUAL.1997.663888

Fahey, M. & McLay, R. (2014). Reproducibility responsibilities in the HPC arena. Reproducibility @ XSEDE Workshop. Retrieved from https://www.xsede.org/documents/659353/703287/xsede14_fahey.pdf

Hampton, S. E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H. (2013). Big Data and the future of ecology. *Frontiers in Ecology and the Environment, 11*(3), 156-162. doi:10.1890/120103

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. *Queue, 10*(2), 30. doi:10.1145/2133416.2146416

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134-140. doi:10.2218/ijdc.v3i1.48

James, D. (2014). Standing Together for Reproducibility in Large-Scale Computing: Report on reproducibility@XSEDE. Retrieved from http://arxiv.org/abs/1412.5557

Jung, C., Gasthuber, M., Giesler, A., Hardt, M., Meyer, J., Rigoll, F., Schwarz, K. Stotzka, R. Streit, A. (2014). Optimization of data life cycles. *Journal of Physics: Conference Series, 513(3).* doi:10.1088/1742-6596/513/3/032047

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization, 10*(4), 271-288. doi:10.1177/1473871611415994

Labrinidis, A., & Jagadish, H. (2012). Challenges and opportunities with Big Data. *Proceedings of the VLDB Endowment, 5*(12), 2032-2033.

Laney, D. (2001) 3D data management: Controlling data volume, velocity and variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larson, E. (1989). They're making a list; Data companies and the pigeonholing of America. The Washington Post (pre-1997 Fulltext), pp. 0-c05.

Lohr, S. (2013). The origins of "Big Data": An etymological detective story. *New York Times*. Retrieved from http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

Lukoianova, T., & Rubin, V.L. (2014). Veracity roadmap: Is Big Data objective, truthful and credible? *Advances in Classification Research Online, 24*(1), 4-15. doi:10.7152/acro.v24i1.14671

Lynch, C. (2008). Big Data: How do your data grow? *Nature, 455*(7209), 28-29. doi:10.1038/455028a

Marx, V. (2013). Biology: The big challenges of Big Data. Nature, 498(7453), 255-260.

Michener, W.K. (2015). Ecological data sharing. *Ecological Informatics, 29*, 33-44. doi:10.1016/j.ecoinf.2015.06.010

Miksa, T., Strodl, S., & Rauber, A. (2014). Process Management Plans. International Journal of Digital Curation, 9(1), 83-97.

Mills, S., Lucas, S., Irakliotis, L., Rappa, M., Carlson, T., & Perlowitz, B. (2012). Demystifying Big Data: A practical guide to transforming the business of government. TechAmerica Foundation, Washington.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Dierolf, U. (2013). Making research data repositories visible: The re3data.org registry. *PLoS One, 8*(11), e78080. doi:10.1371/journal.pone.0078080

Pouchard, L.C., Branstetter, M.L., Cook, R.B., Devarakonda, R., Green, J., Palanisamy, G., . . . Noy, N.F. (2013). A linked science investigation: Enhancing climate change data discovery with semantic technologies. *Earth Science Informatics, 6*(3), 175-185. doi:10.1007/s12145-013-0118-2

Soehner, C., Steeves, C., & Ward, J. (2010). E-science and data support services: A study of ARL Member Institutions. Association of Research Libraries. Retrieved from http://old.arl.org/bm~doc/escience_report2010.pdf

Stodden, V.C. (2010). Reproducible research: Addressing the need for data and code sharing in computational science. *Computing in Science & Engineering, 12*(5), 8-12. Retrieved from http://hdl.handle.net/10022/AC:P:11418

Sukumar, S.R., Natarajan, R., & Ferrell, R.K. (2014). Quality of Big Data in health care. *International Journal of Health Care Quality Assurance, 28*(6), 621-634. doi:10.1108/IJHCQA-07-2014-0080

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One, 6*(6). doi:10.1371/journal.pone.0021101

van Wezel, J., Streit, A., Jung, C., Stotzka, R., Halstenberg, S., Rigoll, F., Garcia, A, Heiss, A., Schwarz, K., Gasthuber, M. (2012). Data life cycle labs: A new concept to support data-intensive science. *arXiv preprint.* arXiv:1212.5596

Ward, J.S., & Barker, A. (2013). Undefined by data: A survey of Big Data Definitions. arXiv preprint. arXiv:1309.5821