A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges

Amir Sinaeepourfard, Xavier Masip-Bruin, Jordi Garcia, and Eva Marín-Tordera
amirs@ac.upc.edu, xmasip@ac.upc.edu, jordig@ac.upc.edu, and eva@ac.upc.edu

Advanced Network Architectures Lab (CRAAX),
Universitat Politècnica de Catalunya (UPC, BarcelonaTech), Spain

Abstract:

Simultaneously to the unstoppable technological evolution and the highly demanding requirements imposed by information exchange in today's world for data sharing, the irruption of novel data collection technologies is driving the need to manage and process notorious volumes of data stored in distributed data repositories all over the world. In addition, the advent of initiatives and technologies define new concepts, such as Big Data, Open Data, and Open Government Data, which enforces the need for data "openness", thus easing data access to main data stakeholders demanding new data usage requirements. Moreover, data production (and collection) is running faster than existing data processing capacity, especially when real-time requirements are set. This scenario turns into a high difficulty and complexity in data handling, including data collection, storing, cleaning, processing, etc., all as a whole identified as the different data life cycles in today's information technology world.

To that end, the paper reviews most relevant Data Lifecycle models, mainly emphasizing on the yet remaining challenges. The paper concludes on two main goals. First, the paper shows that by tailoring Data Lifecycle models to individual needs, there is no global and comprehensive framework, from data creation to data consumption, to be widely utilized in different fields. Second, we go far beyond existing proposals about Vs challenges, proposing a new set of 6Vs challenges that will be used to evaluate these Data Lifecycle models, concluding on the need to work for a widely accepted model matching as much as possible the 6Vs challenges.

Keywords: *Big Data, Open Data, Open Government Data, Data Lifecycle model, Software Lifecycle model, Quality Assurance and Quality Control*

1. Introduction and motivation

Today's Information Technology (IT) world is quickly, continuously and unstoppably progressing empowered by Social Media, Internet of Things (IoT) as well as by emerging smart scenarios, such as smart cities, smart transportation or smart health. This IT progress is driving the need to handle a very huge volume of data as required by the set of services and devices building the envisioned IT smart scenario. As a consequence, a very large amount of data collected in either structured, semi structured or unstructured format (Big Data) are being stored in distributed data repositories to be shared and openly (Open Data) utilized by potential clients for either private or public usage (Open Government Data). However, it is widely accepted that sharing huge amounts of heterogeneous data brings some challenges yet unsolved, mainly related to an efficient and

smart data processing. Some authors have identified the set of challenges that must be addressed, named the 5Vs challenges. These are: Volume, Velocity, Variety, Value and Veracity.

The following three illustrative examples show the growth rate curve for data production. The first example, refers to Big Data, and seats on the Next Generation Sequencing (NGS) [1, 2] in the biological systems field. As shown in Figure 1-1, the data growth evolved from a rate of 1 KB per day in 1996 to a rate of 10 GB per day in 2011. As a result, the size and number of experimental datasets available, from 1996 to 2011, keeps growing exponentially. The second example, refers to Open Data and is promoted by the Government of Catalonia [3, 4]. Figure 1-2 shows the data collected by the app "Mobile Coverage ", starting from January 2014 till now. The data are split into two periods, one lasting 15 months and the other (the more recent) lasting two month. We may easily see how the magnitude of the last one is pretty close to the first, even though the 1 to 15 time relation, what undeniable shows the terrific growth in data. The third example, refers to Open Government Data, and focuses on data offered by three different countries (Australia, United Kingdom and USA) [5-7]. Figure 1-3 shows the datasets growth rate from May 2013 to May 2015. We can easily observe a huge data evolution in both United Kingdom, more than twice from 2013 with 9,440 datasets to 2015 with 24,732 datasets and USA, an impressive 78,8% in the same period, from 73,623 datasets in May 2013 to 131,635 datasets in May 2015.

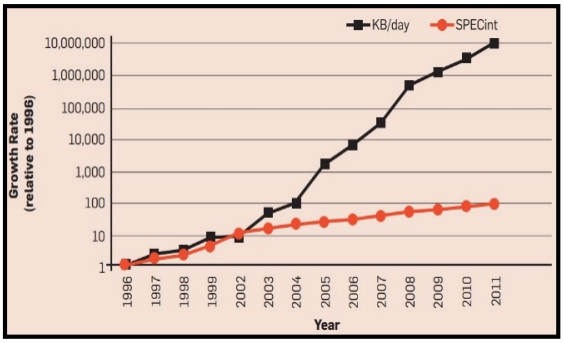


Figure 1-1. Big Data example

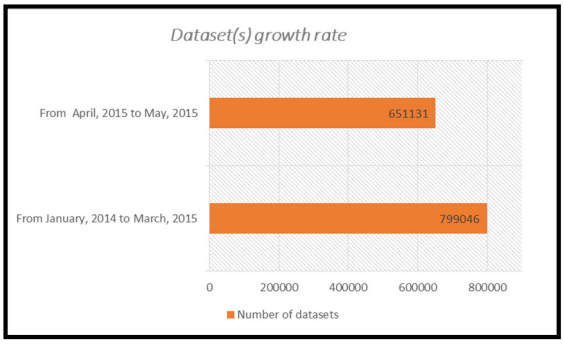


Figure 1-2. Open Data example

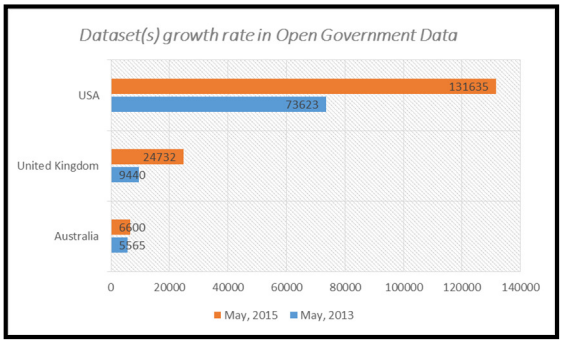


Figure 1-3. Open Government Data example

Figure 1. Some examples of data growth rate from 1996 to 2015

Unfortunately but unavoidably, the continuous increase in the volume and diversity of data are adding high complexity and severe difficulties in all data life stages to be faced by technology stakeholders, particularly when dealing with real time data analytics. Thus, concepts such as Big Data, Open Data and the most recent Open Government Data are stressing the overall data life

cycle processing, while simultaneously easing both users' access to data (leading to transparency, participation, and collaboration for customers [8]) and the deployment of new added-value services. And easing users' access to data and deployment of new services strongly empower the continuous development of the three mentioned concepts, what undeniably sustains and even increases the overall data processing complexity.

For this reason, the worldwide scientific community has invested substantial efforts in the recent decades to overcome the challenges related to managing difficulty and complexity in the different aspects related to data life cycle (e.g. data collection, data processing, data analysis, data storing). To that end, the concept of Data Lifecycle model was formally defined, thus proposing different Data Lifecycle models as a high-level framework encompassing all data management aspects, from data creation to data consumption. The main goals for a Data Lifecycle model are to eliminate waste, to operate efficiently and to prepare data products ready for end-users matching the expected quality constraints [9]. However, Data Lifecycle models are usually tailored to specific fields and interests, turning into particular goals and different data stages depending on the designer's needs for each data stage.

This paper goes deep into the main concepts related to the Data Lifecycle model, aiming at two concrete objectives. First, the paper surveys most of the existing Data Lifecycle models, particularly emphasizing the need for deploying a widely adopted solution not tailored to specific scenarios. Second, the paper points out the weaknesses of existing 5Vs challenges and proposes a novel set of 6Vs challenges to evaluate most relevant Data Lifecycle models, highlighting limitations and weaknesses for each of them. Finally, supported by these two objectives, the paper presents some concluding remarks and open discussions for further actions.

This paper is structured as follows. Section 2 introduces main insights related to Big Data, Open Data, and Open Government Data concepts. Section 3 describes the 6Vs challenges, to be used for evaluation purposes. Then, Section 4 defines the concept of Data Lifecycle model also detailing most of the existing models and after that, Section 5 makes evaluation of the Data Lifecycle models with the respect to the 6Vs challenges. In the Section 6, we start the discussions about the need for a comprehensive global and uniform Data Lifecycle model for data products from data acquisition to data utilization. Finally, section 6 concludes this survey.

2. Big Data, Open Data, and Open Government Data concepts

This section deepens into Big Data, Open Data, and Open Government Data concepts, carefully analyzing pros, cons and main challenges for each initiative. In fact, having a solid knowledge about any ongoing initiative is crucial to get a comprehensive picture about the overall scenario but also to get the required background to facilitate the design and development of innovative solutions fixing the yet unsolved challenges.

2.1 Main Big Data concepts

Coined some years ago, the "Big Data" term [10, 11] has been largely defined by the data scientific community. However, although the technical relevance and business impact of Big Data is widely recognized, there is no global consensus on a uniform and highly accepted definition [12]. That said, after a thorough reading process on the current related literature we may conclude

that Big Data definitions can be categorized into three blocks depending on the main characteristic used to formally establish the definition. The three semantic approaches can be classified as "data size", "technologies and processes", and "challenges". Table 1 highlights relevant references in the literature for each one of the mentioned characteristics:

Table 1. Big Data definitions

Characteristic	***Ref.***	***Description***
Data Size	[13]	Michel Cox and David Ellsworth were among the first to use the term big data literally, referring to the usage of larger volumes of scientific data for visualization.
	[14]	Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
	[15]	Big Data is about the growing challenge organizations face as they deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems.
Technologies and Processes	[12]	Big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.
	[16]	Big Data and data intensive technologies are becoming a new technology trend in science, industry, and business.
	[11]	Big Data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.
	[10]	The process of handling big data encompasses collection, storage, transportation, and exploitation. It is with no doubt that the collection, storage, and transportation stages are necessary precursors for the ultimate goal of exploitation through data analytics, which is the core of big data processing.
	[17]	Big data is a term encompassing different types of complicated and large datasets, all becoming hard to process with the conventional data processing systems.
	[18]	Big data shall mean the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies.
	[19]	Big Data is difficult to process using traditional database and software techniques because a massive volume of both structured and unstructured data that is so large.
Challenges	[11, 20-24]	The "V" discussion is started with 3Vs models. The terms 3Vs were originally introduced by Gartner to describe the elements of big data challenges. So, Big Data is defined as characterized by the three Vs: Volume, Velocity and/or Variety.
	[10]	Later, the 3Vs models were extended to 4Vs. However, there were some doubts on the added V. This reference includes Veracity, so showing the 4Vs as Volume, Velocity, Veracity, and Variety.
	[12]	Another alternative for the added V is value, so defining the 4Vs as Volume, Velocity, Variety, and Value (instead of Veracity).
	[16, 25, 26]	Recently, the discussion moved to 5Vs. Like 4Vs, there are different thoughts to characterize the 5Vs. Some references pushed for a 5Vs model including Volume, Velocity, Variety, Value, and Veracity.
	[17]	This reference pointed out variability (replacing veracity), so defining 5Vs as Volume, Velocity, Variety, Value and Variability.
	[27, 28]	New discussion goes to 7Vs as volume, Variety, Velocity, Veracity, Value, Variability and Visualization.

Big Data provides some opportunities and challenging effects. In short, most opportunities of Big Data refer to its economic impact [29], particularly dealing with optimizing production processes and supply chain, generating new goods and services, targeted marketing, improved organizational management as well as faster research and development. On the other hand, challenging effects of Big Data refer to the challenges highlighted by the exiting 5Vs challenges, mainly summarized into Volume (huge volume of data), Variety (various data formats), Velocity (rapid generation of data), Value (huge value but very low density) [12], and Veracity (quality and security of data) [16] which are later discussed in section 3.

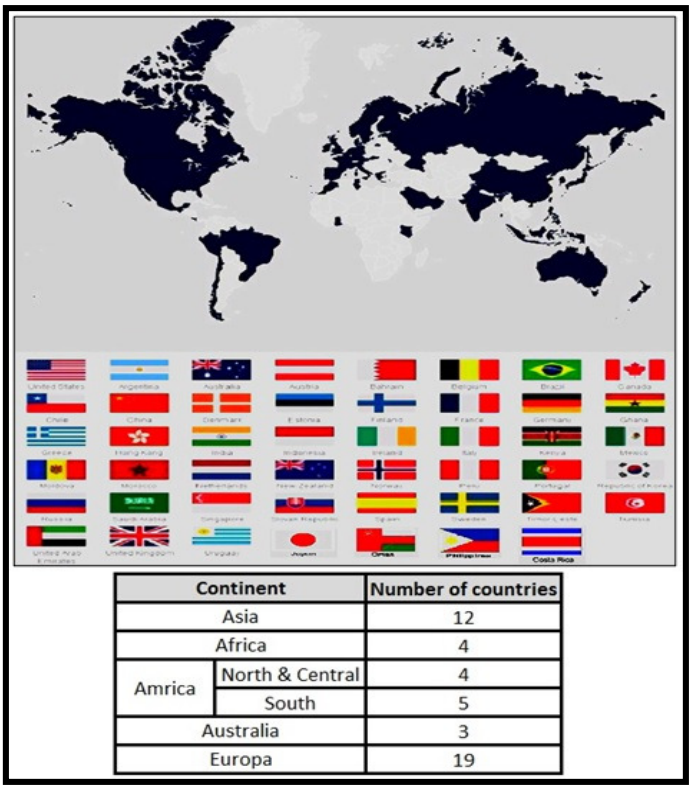
2.2 Open Data

Many profit and nonprofit organizations establish public data spaces, generally referred to as Open Data [8, 29-33], enabling data sharing in a simple and elegant way in public or private spaces. In addition, the data stakeholders send, publish and receive lots of different information through any connectivity technology in place. Thus, data come to open environments with distinct formats and sources by data stakeholders which make new challenges and opportunities for the information world. The main objective of Open Data is to provide a public space for sharing information by data stakeholders. Open Data has some strengths and weaknesses. In short, important benefits in Open Data are: i) business opportunities; ii) a free (or low-cost) public resource fostering innovation and a better-informed public [29]; iii) capacity to generate more services for users, and; iv) a more vibrant economy [31]. On the other side, Open Data has also some weaknesses, such as: i) the lack of data quality; ii) incompatible formats and access methods, and; iii) various semantic interpretations of data [32].

2.3 Open Government Data (Open-Data Government)

Several initiatives have been recently set by many governments around the world to create Open Government Data portals. Open Government Data can be considered as Open Data provided by governments. This means that the data provided must have a valid and trusted reference for users. Therefore, this environment provides such a reliable source, with an acceptable data quality, for customers. The aim of Open Government Data is to connect the customers to the trustable and reliable information for taking some advantages of better services. Its main rationale can be decoupled into a better governance, great and improved services, and a more vibrant economy [30, 34-39]. Forty-seven countries [40], as of May 2015, are already participating in the Open Government Data model throughout cities, states, and countries as shown in Figure 2.

Open Government Data portals bring huge pros and cons for Open Government Data stakeholders. Plus, those pros and cons can be added to Open Data strengths and weaknesses which have been discussed in last part. On one side, major advantages of Open Government Data are transparency, participation, and collaboration [8], and valid sources and standardization of data format. On the other side, the most notable barriers are the risk of violating legislation by opening data, difficulties with data ownership, misinterpretation and misuse of raw data, negative consequences of transparency and negative consequences for the government [41].



Continent		Number of countries
Asia		12
Africa		4
Amrica	North & Central	4
	South	5
Australia		3
Europa		19

Figure 2. Countries joining the Open Government Data portals

3. Extending the 5Vs challenges to a 6Vs model

The main challenges in Big Data have traditionally been described through the 3Vs challenges, Volume, Variety and Velocity, as defined by Gartner [11]. This model has been extended to the 5Vs challenges, which may be considered as 4 +1, since the latter differs depending on the reference. Indeed, 4Vs parameters include Value, Variety, Velocity and volume. The additional one may be either Variability as stated by [17, 42] or Veracity as read in [16, 25, 26]. Recently, there is some effort to show that the challenges can assume 7Vs [27, 28], including both Variability and Veracity, and adding Visualization as a new challenge. And some other authors propose Volatility, Viscosity and Virality as additional main challenges [43-57]; however, we think these are not mature enough to be considered in this work. Next we introduce main references and concepts for all previous challenges among all sources.

1- Value of data

Value is a highest priority aspect of Big Data [11]. This is rooted on the fact that the goal for data analysis and management, indeed, is to obtain enriched information. To reach this goal, we must explore large amounts of data with different data formats and sources to pick up some hidden data which can build the valuable information for business and end-users purposes [12]. The main challenge linked to this parameter refers to provide smartness approaches and scenarios for discovering and recognizing hidden value of information among all data.

2- Volume of data

This parameter indicates the huge amount of data which can be produced by different data sources with distinct data types and formats and must be managed somehow. Note that Big Data concepts refer to the fact that a very large amounts of data must be managed and analyzed. Furthermore, traditional technologies can efficiently analyze and manage datasets limited to a certain size typically in the size of Gigabyte. Therefore, the challenge is to promote that new techniques and technologies to handle these amounts of massive data, with different formats and sources, as well [10, 12].

3- Variety of data

Variety refers to the fact that data comes from different sources, like sensors, social networks, smartphones, and so on [11] and, therefore, the formats to be considered may be very diverse. These includes structured, semi-structured, and unstructured data, such as audio, video, webpage, plain text, etc. [12]. In addition, the continuous technology progress will definitely bring new devices and solutions enabling additional data collection (for example, we may envision highly impacting advances in the e-health or in the transportation sectors), what is also adding variability to the collected data. Thus, handling data heterogeneity is a relevant challenge for big data. Indeed, traditional tools, such as SQL, use to handle only structured databases, but do not perform well when the databases are semi-structured or unstructured [58].

4- Velocity of data

Velocity highlights that the speed of the data stream generates extremely quick; for instance, sensors produce streams of data very fast, and the number of sensors are uncounted in the smart scenarios. This demands a link between Big Data concepts, specifically data collection, processing, analysis and so on, and business value timely and efficiently for obtaining the expected value [10, 12, 58]. In addition, the business markets make plans for having more frequent decision making and being closer to their customers' requirements. For instance, the bank industry needs to get online, or nearly online, data analytics, which it goes to Fast Data analytics concepts, for creating better services for their customers in these days. Traditional algorithms and systems cannot manage the current data stream growth rate nor can process the increasingly growing data sizes [10, 23]. So, the challenges are to provide appropriate technologies to satisfy the business requirements.

5- Variability of data

Variability refers to the fact that data meanings can be changing and updating over the time. It refers to data semantic concepts which are related to the intrinsic and interpretations meanings of data [28]. For instance, sometimes one single word translates to multiple meanings. Or one word can be translated with the different meaning depending on the sentence context. Or even some words can change to different meanings throughout the time [27, 28]. This parameter is highly impacting stakeholders involved in data analysis [59]. The challenge points that we need to define the specific algorithms and approaches, like sentiment analysis and opinion mining, which will make text deeply and globally understandable [60].

6- Veracity of data

Data veracity can be seen from two different points of view, namely quality concepts and security concepts, both defined next.

Veracity in Big Data, from a security perspective, guarantees that the data access will be secure, that is, unauthorized access and modification will be prevented. This makes data to be trusted, authentic and protected for end-usage [16]. The challenge is to guarantee that huge sets of data will be preserved against any unexpected change and attack during collection, processing, storing and any other stage during the whole data lifecycle [16].

Veracity in Big Data, from a quality perspective, guarantees that the data provided are the best fit for use by end-users [61]. The issue of data quality has been considered by a number of researchers, and includes topics such as data complexity, missing values, noise, imbalance, dataset shift and so on [62-66], and concentrate on details about how the data can be relied for making the best customers' decisions among all data [50, 67]. The challenge faces that despite the abundant data being available for usage, the quality of the data could be too complex for decision making [10]. Furthermore, data quality can be guaranteed with two different strategies that try to prevent or correct errors throughout any activity [68, 69]:

- Quality Assurance (QA) is "a part of quality management focused on providing confidence that quality requirements will be fulfilled" regarding the ISO 9000 standard definition [70]. QA tries to prevent any kind of defects in the product with a focus on the process which is used to build the product [71].
- Quality Control (QC) is "a part of quality management focused on fulfilling quality requirements" regarding the ISO 9000 standard definition [70]. QC tries to identify and correct defects in the final products [68, 71].

7- Visualization of data

Data visualization refers to the way of presenting the data, once it has been processed, into something easily visible, readable, understandable and tangible for most of the audiences like tables, diagrams, images and other intuitive display techniques[27]. The visualization may help users to perform analysis. This analysis can give some possible solutions for adding better quality and performance to the business. The challenge highlights that Big Data visualization cannot be assumed and worked easily like traditional datasets. Some challenging examples in Big Data visualization are visual noise, information loss, large image perception, high rate of image change and high performance requirements) [72].

After reviewing all definition about challenges of the Big Data concepts, we believe that there is a difference between using the Vs model for Big Data definition and Big Data challenges. The complete definition of Big Data can be generated with variety, volume, and velocity, because these are the features that describe Big Data, and perhaps value, because Big Data technology has appeared due to the potential value among such massive data.

With respect to the Big Data challenges, we do not believe that visualization, referring to a way of presenting data once it's been processed [26], is one main challenge for Big Data technology

because it is an optional software programming aspect for end-users. In this paper we propose the 6Vs challenges model, which includes Value, Volume, Variety, Velocity, Variability and Veracity, as a model to evaluate the comprehensiveness of the different data lifecycle models considered in section 4.

Finally, this paper proposes that 6Vs challenges can be assumed for Open Data and Open Government Data concepts as well because Open Data and Open Government Data face with the same challenges. However, the density of those challenges are different in Big Data, Open Data, and Open Data Government concepts. For instance, Open Government Data is less challenging in terms of Veracity because all datasets have been prepared with the unified sources, but there are still some major challenges in terms of security and data quality under the Veracity challenge [41].

4. Data Lifecycle model

In a broad range of data areas, many researchers in academia and industry propose Data Lifecycle models to manage data life as valued assets [73-76]. Data Lifecycle models define a high level framework which is a global view of data life from production stage to consumption stage. The goal of a Data Lifecycle model is to optimize data management, from efficient organization to elimination of any kind of waste, in order to provide data products appropriate for use by end-users matching the expected quality requirements [9, 61]. One simple Data Lifecycle model could be represented with three elements which are Acquisition, Curation, and Preservation. [76]. Acquisition refer to the specific process that can collect the information. Curation means to manage the data from creation, like an initial feed, which can be used for final utilization, such as discovery and re-use, in the future. Preservation is to keep available the data in any kind of physical sources for future usage by special processes.

In this section we review most Data Lifecycle models found in the literature. Some models have been proposed to manage data for a specific scientific field or project and, therefore, these models just consider some elements of a complete lifecycle of data. And some other models concentrate on specific data phases, such as data curation or data preservation. For this reason, in this work we evaluate the completeness of each model with respect to the previously defines 6Vs challenges model.

4.1. The ANDS Data Sharing Verbs model

ANDS is abbreviation of Australian National Data Service. ANDS put together information from data providers and publicly invested institutions, to be used by research institutions within Australia [77-79]. The ANDS Data Sharing Verbs model aims to design systems for supporting data sharing and re-using. The model offers Create, Store, Describe, Identify, Register, Discover, Access, and Exploit as steps of Data Lifecycle model [77, 80-82]. So this data model solves enquires to exchange information for Australian research with simple access possibilities through an internet based discovery [78].

Challenge: This model cannot be considered as a comprehensive model because it was created for a specific purpose which is data sharing and data re-use. In addition, there is not any focus on data quality, neither QA nor QC.

4.2. The BLM model

The BLM (Bureau of Land Management) administrates the public lands in the USA [83]. They propose a model for sharing information among customers and providing high level of quality. Plan, Acquire, Maintain, Access, Evaluate, and Archive are steps of the BLM model designed for land data management [84]. This model has been designed as a non-linear representation, where QA and QC management are central issues. So this model seems to work for data archiving and accessing with QA and QC management, as depicted in Figure 3 [85].

Challenge: It cannot be assumed as a comprehensive model because the orientation of the model is for exchange information with emphasis in data quality. In addition, this model applies for a particular field, related to public landing information. Further, it can be discussed more about time and cost efficiencies in this non-linear model because quality assurance and quality control are in the center of this model.

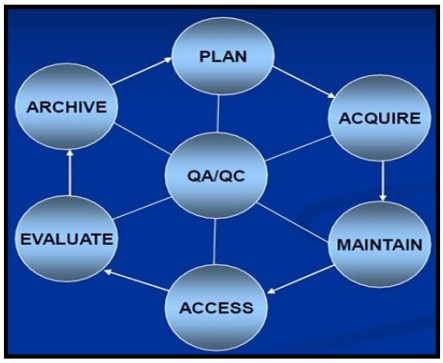


Figure 3. The BLM model

4.3. The CSA model

The Cloud Security Alliance (CSA) is the world's leading organization to manage secure cloud computing environments [86]. The CSA proposes a data lifecycle model for data security in the cloud environment. The data model provided has six phases which are Create, Store, Use, Share, Archive, and Destroy, as seen in Figure 4 [87]. So this model addresses one particular problem, security, in the cloud computing environment.

Challenge: This model cannot be a comprehensive data model because it has been designed for data security in the cloud computing models. Therefore, concepts such as data quality, data processing and data analysis are not considered.

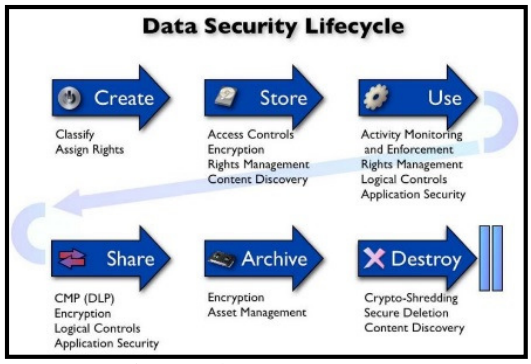


Figure 4. The CSA model

4.4. The DataONE model

The Data Observation Network for Earth is an organization, called DataONE, which is funded by the US National Science Foundation (NSF) [88, 89]. Their data model aims to provide data preservation and re-use for research in biological and environmental sciences. The proposed data life cycle includes Collect, Assure, Describe, Deposit, Preserve, Discover, Integrate, and Analysis, as illustrated in Figure 5 [89-91]. So this model can be used for storing and retrieving information for long term usage.

Challenge: This model has been developed specifically for data preservation and re-use, which cannot be seen as a comprehensive model. In addition, there is not any focus about data security.

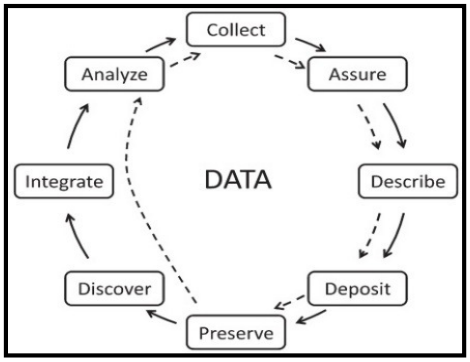


Figure 5. The DataONE model

4.5. The DCC model

The Digital Curation Centre (DCC) is an organization that works for the digital information curation to improve the higher education in the United Kingdom [92-95]. The DCC provides a model for successful curation and preservation of data, where data is in digital form [96]. The DCC lifecycle includes different layers: Full Lifecycle Actions, Sequential Actions, and Occasional Actions. Full Lifecycle Actions are divided into four steps which are Description and Representation of Information, Preservation Planning, Community Watch and Participation, and Curate and Preserve. Sequential Actions provide seven steps which are Conceptualize, Create or Receive, Appraise and Select, Ingest, Preservation Action, Store, Access, Use and Reuse, and Transform. Occasional Actions include Dispose, Reappraise and Migrate, as shown in Figure 6 [85, 92-94, 96-99]. The steps disposition in this model is quite sophisticated as they are placed in a multiple layer cyclic structure. This model has been designed for successful curation and preservation of digital data.

Challenge: This model cannot be labeled as a comprehensive model because it has been designed specifically for curation and preservation of data. Plus, data analysis and data integration are not considered in this model. In addition, this model does not provide QA because the "Appraise and Select" step works like QC of data in the lifecycle [92].

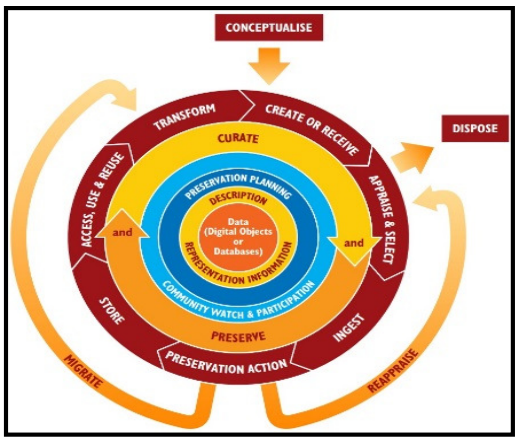


Figure 6. The DCC model

4.6. The DDI conceptual model, version 3.0

The Data Documentation Initiative (DDI) is one project of the Inter-University Consortium for Political and Social Research (ICPSR) [77, 100]. The DDI tries to produce a metadata specification for the description of social science data resources [100]. The offered model includes eight elements, as shown in Figure 7, which are Study Concept, Data Collection, Data Processing, Data Archiving, Data Distribution, Data Discovery, Data Analysis, and Repurposing [77, 81]. This model generates a conceptual model for political and social data research and standardization. In version 3.0 they provide the standardization for XML vocabularies [101].

Challenge: This is almost a comprehensive model because they address successfully most steps in the data lifecycle, from collection to consumption. However, it seems there is not any focus on data quality and data security.

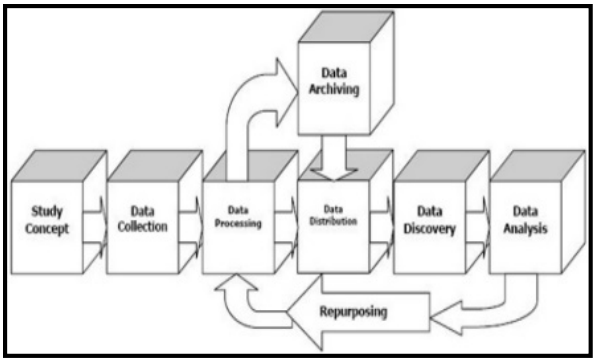


Figure 7. The DDI Conceptual model, Version 3.0

4.7. The DigitalNZ Content model

DigitalNZ comes from Digital New Zealand. Their goal is to collect and increase the amount of digital content for users, and the data model is designed for archiving and using the digital information [102, 103]. The proposed model includes Selecting, Creating, Describing, Managing, Preserving, Discovering, and Using and Reusing as steps, as shown in Figure 8 [104]. This model aims to manage digital information exchange among data stakeholders.

Challenge: This model has been designed to focus only on archiving and using purposes, so it cannot be considered a comprehensive model. Plus, data analysis, data integration, data security and data quality is not provided in this model.

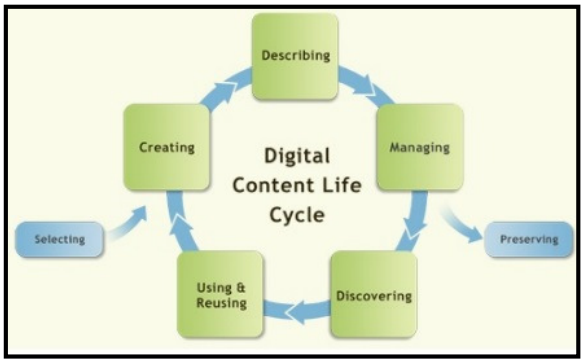


Figure 8. The DigitalNZ Content model

4.8. The Ecoinformatics model

Ecoinformatics is a framework to help scientists working with the relevant biological, environmental and socioeconomic data and information. The data model aims to build new knowledge through creative tools and approaches for discovering, managing, integrating, analyzing, visualizing and preserving relevant data and information [73, 77]. As depicted in Figure 9, Plan, Collect, Assure, Describe, Preserve, Discover, Integrate, and Analyze are steps of this model [77, 105]. So the model provides a framework to achieve new insights of data and information for some particular sciences.

Challenge: This framework design is almost a comprehensive model because this has been designed for data collection, data preservation, data discovery, and some data manipulation, such as data integration and data analyze. However, data security is still an open challenge. So this is the reason why it is not actually a comprehensive model. This model looks very similar to the DataONE model described in Section 4.4 but they differ on the first step.

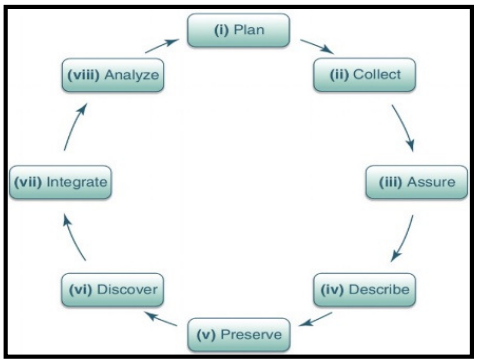


Figure 9. The Ecoinformatics model

4.9. The Generic Science model

The Generic Science model is offered by the Science Agency to manage scientific digital data [106]. This model can be useful to manage data collection methods for archiving or disposing data. The Generic Science data model has Plan, Collect, Integrate and Transform, Publish, Discover and

Inform, and Archive or Discard as six stages of the lifecycle [102]. This model, shown in Figure 10, can predict the next set of data acquisitions with specific techniques to use for data management plans [102].

Challenge: This model is not a comprehensive model for the whole data life cycle because it has been designed specifically for data archiving and disposing. This model in not concerned about data analysis, data security and data quality in the cycles.

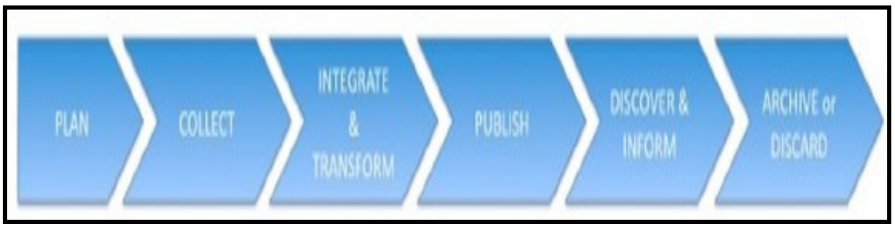


Figure 10. The Generic Science Model

4.10. The Geospatial model

The Geospatial Data Lifecycle model is supported by the Federal Geographic Data Committee (FGDC) [107, 108]. The model aims to explore and save valuable information for the geographic and related spatial data activities. Figure 11 summarizes the Geospatial Data Lifecycle stages which are Define, Inventory/Evaluate, Obtain, Access, Maintain, Use/Evaluate, and Archive [85, 107]. This model is handled to discover data with acceptable quality and business requirements for future use.

Challenge: This model cannot be used as a comprehensive model because it has been designed specifically for searching and archiving information. Plus, this model does not address anything about data analysis and data integration in the cycle. Furthermore, QA and QC are included in each stage which can be a limitation for time and work efficiency in this model.

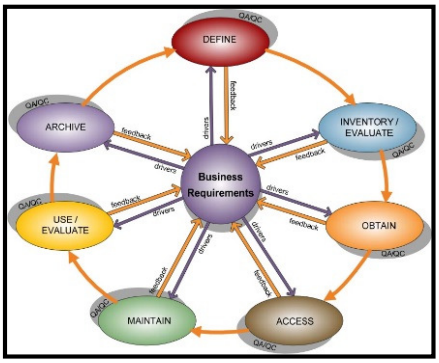


Figure 11. The GeoSpatial model

4.11. The LOD2 Stack model

LOD2, the Linked Open Data, is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Program [109]. The LOD2 Stack data model searches useful data which can be fitted to the end-user requirements. This model includes Storage/Querying, Manual revision/Authoring, Interlinking/Fusing, Classification/Enrichment, Quality Analysis, Evaluation/Repair, Search/Browsing/Exploration,

and Extraction as the different phases, as shown in Figure 12 [77, 110, 111]. This model is helpful to find relevant data for end-users.

Challenge: This model cannot be assumed as a comprehensive model because it concentrates on searching some desirable data. In addition, they do not address data security. The LOD2 Stack model manages partially quality, but only QC, as they just measure the quality of the web contents [110].

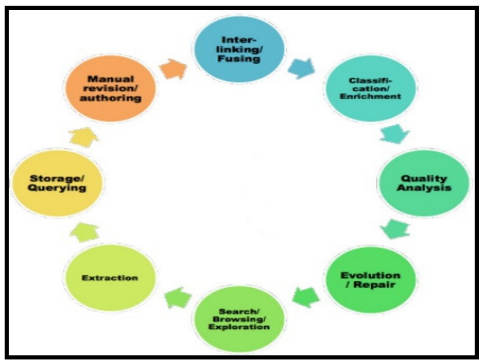


Figure 12. LOD2 Stack Model

4.12. The University of Deusto model

A group of researchers from the University of Deusto, in Spain, have proposed one Data Lifecycle model for data management in smart cities [77]. As depicted in Figure 13, the different stages of this model are Discovery, Capture, Curate, Store, Publish, Linkage, Exploit and Visualize. This model is an option to apply for discovering, storing, and publishing data in smart cities.

Challenge: The model cannot be considered as a comprehensive model because it is specific for data management in smart cities. In addition, there is not any focus on data security, nor on data quality (including QA and QC) in the model.

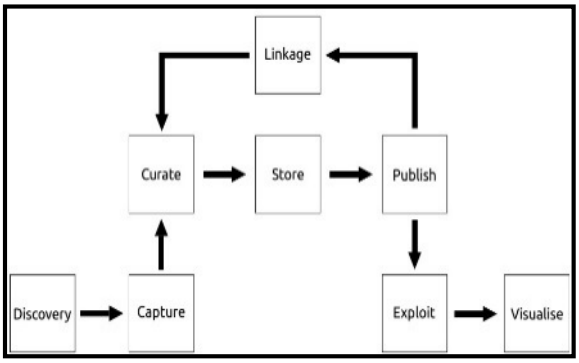


Figure 13. The University of Deusto model

4.13. The Records model

The University Archives and Historical Collections (UAHC) department assists the Michigan State University to do the efficient administration and management under university's procedures [112]. The Records data model aims to offer a solution for moving paper work to digital work in any kind of offices, especially in the university. The offered model includes Create/Receive, Use and File, Transform and Store, Dispose and Archive/Destroy, as different steps shown in Figure

14 [85]. The model provides an electronic procedure for making more efficient and better administration and management in the university.

Challenge: This model cannot be considered as a comprehensive model because it focuses on data archiving. Plus, it does not include any concept about data quality, data analysis, data processing, and data integration.

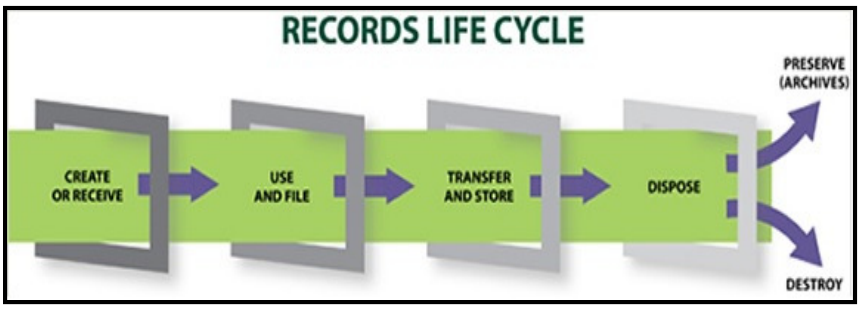


Figure 14. The Records model

4.14. The JISC Research model

The Joint Information Systems Committee (JISC) under the Managing Research Data Programme, works for good research data management and sharing for the UK Higher Education and Research [113]. The Research model proposed has been designed for sharing data among users. The model includes seven steps: Plan, Create, Use, Appraise, Publish, Discover, and Reuse, as seen in Figure 15 [102, 114]. The offered model is a framework for data sharing and discovery as a part of their global data management initiatite.

Challenge: This model cannot be used as a comprehensive model because it has been designed for data sharing and discovery. This model does not offer any stage for data processing, data integration and data analysis. Plus, this model covers QC concepts under "Appraise" step but QA is not provided.

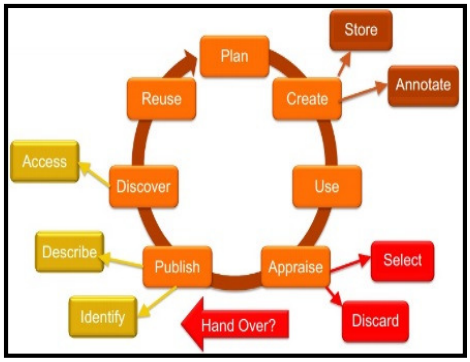


Figure 15. The JISC Research model

4.15. The UK Data Archive model

The UK Data Archive works among the largest collection of digital data, including social and economic data, in the United Kingdom [115]. 4.15. The UK Data Archive model focuses on acquisition, curation and archive of the digital data. The model has Creating Data, Processing Data, Analyzing Data, Preserving Data, Giving Access to Data, and Re-using Data, and organize them

as a cycle, as shown in Figure 16 [77, 81]. So the model can be a good choice for archiving and discovering across the digital data.

Challenge: This model can be assume as a comprehensive model because it provides the full data lifecycle, which includes acquisition, curation and preservation. However, the model concentrates on particular social and economic sciences. However, data quality is not covered in this model.

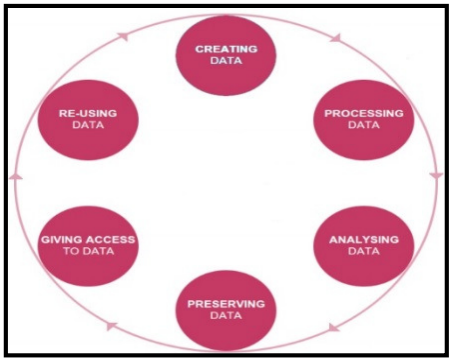


Figure 16. The UK Data Archive model

4.16. USGS model

The U.S. Geological Survey (USGS) Community for Data Integration (CDI) works with data and information management issues that can be relevant for the U.S. Bureau's Scientific Research [116]. The USGS data model provides a framework to evaluate and improve policies and practices for managing the scientific data, and to identify areas in which new tools and standards are needed [9, 116]. The model includes Primary and Cross-Cutting model elements, as depicted in Figure 17 [9, 85, 116]. The Primary model elements are Plan, Acquire, Process, Analyze, Preserve, and Publish/Share. Besides that, Cross-Cutting model elements comes with Describe, Manage Quality, and Backup and Secure, as steps. Thus, this model can be a reference to manage the scientific data for having better standards and tools.

Challenge: This model can be considered a comprehensive model because it suggests data cycles for acquisition, curation and preservation. However, this model chooses a linear presentation for the graphic model, so time and work efficiency should be under discussion, especially for large amounts of data. In addition, the model does not cover data security because the meaning of secure in the "store and secure" element refers to physical risk, such as hardware and software failures [116].

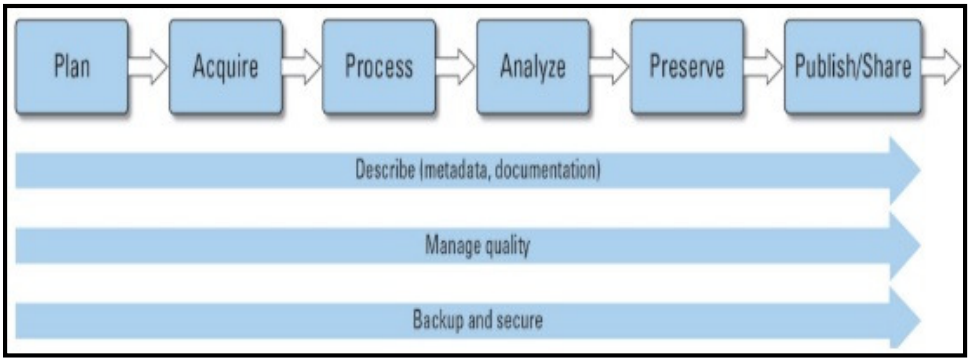


Figure 17. The USGS model

4.17. The Beijing University model

The model comes under one research group from the Beijing University on Posts and Telecommunication, in China. This model is used for data security in the cloud computing environment. The graph-based model, depicted in Figure 18, has five stages which are Create, Store, Use and Share, Archive, and Destruct [117]. This model is appropriate for the security in the cloud environments.

Challenge: This model cannot be considered as a comprehensive model because it is designed only to support data security in the cloud. Plus, it does not cover data quality, data analysis and data publishing in any of the stages.

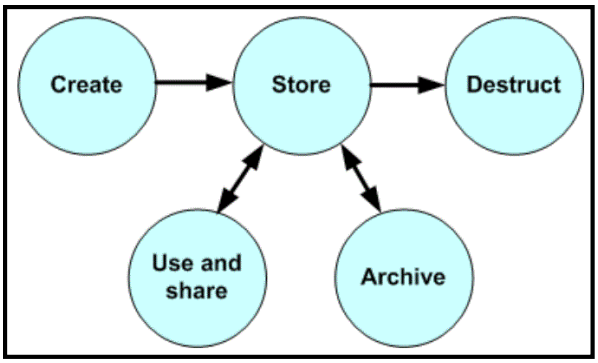


Figure 18. The Beijing University model

5. Evaluation of the Data Lifecycle models

We have analyzed and described most Data Lifecycle models found in the literature. The advent of Data Lifecycle models have depicted that the new requirements about data management and mobility are adding to the traditional Data Lifecycle models some specific steps, such as data quality, data security, business focus, and so on. In addition, many models have been designed tailored to the specific purpose to solve one particular problem or area in science(s) and/or data management. Of course, it is obvious that each model is a suitable design regarding their research or project requirements and, perhaps, they leave some challenges to address because these are out of the scope of their objectives. For this reason, we wonder if there is a global and comprehensive Data Lifecycle model that can successfully deal with most researchers and projects requirements. The main contribution of a comprehensive model is to eliminate waste and duplicity in researchers' task to design a new model for any new project. In this section we evaluate all introduced models with respect to the 6Vs challenges, including Value, Volume, Variety, Velocity, Variability, and Veracity, in order to know to what extent each model is comprehensive.

Table 2 shows the results of the Data Lifecycle models evaluation, with respect to the 6Vs challenges. The evaluation is marked "yes" (□) at each box if the model can address and handle the corresponding V challenge or, otherwise, it is marked "no" (x) at each box if the model cannot manage the V challenge.

Table 2: Make evaluation of the Data Lifecycle models

Data Lifecycle models		6Vs Challenges							
		Value	Volume	Variety	Velocity	Variability	Veracity		
							QA	QC	Security
1	ANDS Data Sharing Verbs model	x	☑	☑	x	x	x	x	☑
2	BLM model	☑	☑	☑	x	x	☑	☑	☑
3	CSA model	x	☑	☑	x	x	x	x	☑
4	DataONE model	x	☑	☑	x	☑	☑	☑	x
5	DCC model	☑	☑	☑	x	x	x	☑	☑
6	DDI conceptual model, version 3.0	☑	☑	☑	☑	☑	x	x	x
7	DigitalNZ Content model	x	☑	☑	x	x	x	x	x
8	Ecoinformatics model	☑	☑	☑	x	☑	☑	☑	x
9	Generic Science model	☑	☑	☑	x	x	x	x	x
10	Geospatial model	☑	☑	☑	x	x	☑	☑	☑
11	LOD2 Stack model	x	☑	☑	☑	☑	x	☑	x
12	University of Deusto model	x	☑	☑	☑	☑	x	x	x
13	Records model	x	☑	☑	x	x	x	x	x
14	JISC Research model	☑	☑	☑	x	x	x	☑	☑
15	UK Data Archive model	x	☑	☑	☑	☑	x	x	☑
16	USGS model	☑	☑	☑	☑	☑	☑	☑	x
17	Beijing University model	x	☑	☑	x	x	x	x	☑

In order to complete this table, we have assumed that, as volume and variety are fundamental challenges in Big Data management, all Data Lifecycle models are able to address them. For that reason, we have marked “yes” in the table for all models. In addition, most Data Lifecycle models have been designed to manage data in specific environments, so this suggests some challenges that have certainly been addressed. For example, the CSA model has been proposed to provide security for Cloud computing environments. Therefore, this model is appropriate to be considered in environments where volume and variety of data has to be managed, so we mark “yes” for volume and variety. But value, velocity, variability, and veracity must be deeper reviewed. Eventually, we have marked “yes” in security, as part of veracity, because this model is specific for data security, but we have marked “no” for variability, velocity, and QA and QC, as another part of veracity, because in their description they do not show any focus on those challenges.

From the Table we can observe that data quality, as part of veracity, and velocity, are challenges that have only been considered in very few models. This means that a comprehensive Data Lifecycle model must pay more attention to these challenges, so guaranteeing data quality and fast data generation are some important keys in Big Data management. Finally, the table shows that there is not any Data Lifecycle model that covers completely all the 6Vs challenges among their lifecycle phases. The USGS model get closer to this completeness; however, there are still some

lack in data security in this model. Therefore, we conclude that there does not exist any whole global and comprehensive model with respect to the 6Vs challenges in this evaluation.

6. Discussions and future directions

In this section, we highlight the results of our evaluation and propose some future work in order to organize data management and processing more efficiently. We have organized these conclusions in three main blocks:

1- Strong support to the Data lifecycle Model

Data is an important resource. In fact, sometimes data is considered the new kind of oil by many communities in this recent century [116, 118, 119]. Therefore, it is necessary to manage data efficiently, from creation to consumption. A Data Lifecycle model is a well-designed framework designed to organize data products and, as has been shown in this paper, there are many efforts to propose new frameworks to manage data through Data Lifecycle models. Furthermore, some recent efforts are moving to new scenarios that make connections between Data Lifecycle models and Software Lifecycle models to generate more efficient data products for customers in any kind of sciences [75, 120]. Indeed, we believe that attention has to be paid for data movement and management in any kind of scenario to eliminate any chance for future problems.

2- Standardization and globalization of the Data Lifecycle model

As shown in the Data Lifecycle models review in Section 4, there exist many Data Lifecycle models as part of science(s) or project(s) in the academia and industries nowadays. Each group has invested lots of time and efforts to design their own Data Lifecycle model in their project or science. In addition, this view of particular data management, could provide some concern in future related to data integration, data quality, and so on. For instance, if data integration should be done in the future, different levels of the data quality policies could generate conflicts, or provide unusable and useless data and information in the data repositories. So we propose to define standardization and globalization for the Data Lifecycle models, in order to prevent any kind of limitations in future data management work.

3- Designing a comprehensive Data Lifecycle model

A comprehensive model can provide a global framework with a complete view that can be tailored to any science and project that may consider all current and future technical issues. The model can help designers to grab their requirements faster than the current way. The advantage of a comprehensive model is to provide a well-designed model with time and work efficiencies, that eliminates waste and duplicities in researchers and projects design. Furthermore, the model should be designed flexible to support and adapt to any kind of challenges and technologies which can be added in the future. So we propose that it is necessary to design a comprehensive Data Lifecycle model, valid for most research groups.

7. Conclusions

There is no doubt that Data is a valuable asset in the recent decades. So, we must use and reuse this asset for taking some advantages in our daily business and life as much as possible by the efficient and smartness solutions. Plus, the new technologies are ascertained many new challenges

and complexities for data life movement from creation to consumption in terms of different data formats and a very large amount data. Big Data, Open Data and Open Government Data are some new concepts to tailor with the new technologies and innovations concepts, which define new challenges as Vs Challenges.

In this paper, we have surveyed most existing Data Lifecycle models which are a high level solution to manage data life from production to usage. The objective of this paper is to introduce a novel 6Vs challenges for the massive amount of data. In addition, we use the 6Vs challenges to make an evaluation of the described Data Lifecycle models. Eventually, the result of this evaluation show that there is not any comprehensive Data Lifecycle model to manage the data life from first to end in today's data world. We delight to highlight that a comprehensive Data Lifecycle model can provide several profits. The main profit is to help saving more time and efforts for the researchers in the academia and industries, having a high level of work efficiency for data management and offering more flexibility to join any new ideas and technologies to this proposed model.

Acknowledgments

The authors would like to acknowledge the support received from the Catalonia Government, Generalitat de Catalunya, under FI-DGR contract.

References

1. Jagadish, H., et al., *Big data and its technical challenges.* Communications of the ACM, 2014. **57**(7): p. 86-94.
2. Kahn, S.D., *On the future of genomic data.* Science(Washington), 2011. **331**(6018): p. 728-729.
3. gencat), O.p.d.o.t.g.o.C.O.D. *Data collected by the app*. 2015; Available from: http://dadesobertes.gencat.cat/en/cercador/detall-cataleg/?id=7710.
4. Catalunya, G.d. *Cobertura mòbil Application*. Available from: http://coberturamobil.gencat.cat/web/index_en.
5. *Open Government Data in Australia*. Available from: http://www.data.gov.au/.
6. *Open Government Data in United Kingdom*. Available from: http://data.gov.uk/data/search.
7. *Open Government Data in USA*. Available from: http://catalog.data.gov/dataset.
8. Lassinantti, J., *Public Sector Open Data.*
9. USGS. *USGS Data Lifecycle Overview*. Available from: http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php.
10. Zhi-Hua, Z., et al., *Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives [Discussion Forum].* Computational Intelligence Magazine, IEEE, 2014. **9**(4): p. 62-74.
11. Hashem, I.A.T., et al., *The rise of "big data" on cloud computing: Review and open research issues.* Information Systems, 2015. **47**: p. 98-115.
12. Chen, M., S. Mao, and Y. Liu, *Big data: A survey.* Mobile Networks and Applications, 2014. **19**(2): p. 171-209.
13. Cox, M. and D. Ellsworth. *Managing big data for scientific visualization*. in *ACM Siggraph*. 1997.
14. Manyika, J., et al., *Big data: The next frontier for innovation, competition, and productivity.* 2011.

15. Villars, R.L., C.W. Olofson, and M. Eastwood, *Big data: What it is and why you should care.* White Paper, IDC, 2011.
16. Gordijenko, D. *Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure*. in *Secure Data Management: 10th VLDB Workshop, SDM 2013, Trento, Italy, August 30, 2013, Proceedings*. 2014. Springer.
17. Samuel, S.J., et al., *A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES.* 2006.
18. Micheni, E.M., *Diffusion of Big Data and Analytics in Developing Countries.* 2015.
19. Bloomberg, J., *The Big Data Long Tail*. 2013.
20. Zikopoulos, P., et al., *Harness the Power of Big Data The IBM Big Data Platform*. 2012: McGraw Hill Professional.
21. Berman, J.J., *Principles of big data: preparing, sharing, and analyzing complex information*. 2013: Newnes.
22. Laney, D., *3D data management: Controlling data volume, velocity and variety.* META Group Research Note, 2001. **6**: p. 70.
23. O'Leary, D.E., *Artificial intelligence and big data.* IEEE Intelligent Systems, 2013. **28**(2): p. 0096-99.
24. Gartner, I. *Big Data*. Available from: http://www.gartner.com/it-glossary/big-data.
25. Rossi, R. and K. Hirama, *Characterizing Big Data Management.* Issues in Informing Science and Information Technology, 2015. **12**.
26. Narasimhan, R. and T. Bhuvaneshwari, *Big Data–A Brief Study.*
27. MCNULTY, E. *Understanding Big Data: The Seven V's*. 2014; Available from: http://dataconomy.com/seven-vs-big-data/.
28. *Understanding the 7 V's of Big Data*. 2015; Available from: http://www.optimusinfo.com/blog/understanding-the-7-vs-of-big-data/.
29. GURIN, J., *DRIVING INNOVATION WITH OPEN DATA.* THE FUTURE OF, 2014: p. 55.
30. Ridgway, J. and A. Smith. *Open data, official statistics and statistics education: threats, and opportunities for collaboration*. in *Proceedings of the Joint IASEIAOS Satellite Conference "Statistics Education for Progress", Macao, China*. 2013.
31. Ren, G.-J. and S. Glissmann. *Identifying information assets for open data: the role of business architecture and information quality*. in *Commerce and Enterprise Computing (CEC), 2012 IEEE 14th International Conference on*. 2012. IEEE.
32. Masip-Bruin, X., et al. *Unlocking the Value of Open Data with a Process-Based Information Platform*. in *Business Informatics (CBI), 2013 IEEE 15th Conference on*. 2013. IEEE.
33. Gurin, J., *Open data now: the secret to hot startups, smart investing, savvy marketing, and fast innovation*. 2014: McGraw Hill Education.
34. Fox, M.S., *City data: Big, open and linked.* Department of Mechanical and Industrial Engineering University of Toronto, 2013.
35. Al-Khouri, A.M., *Open Data: A Paradigm Shift in the Heart of Government.* Journal of Public Administration and Governance, 2014. **4**(3): p. Pages 217-244.
36. Jetzek, T., M. Avital, and N. Bjørn-Andersen, *The Value of Open Government Data.* Geoforum Perspektiv, 2013. **23**: p. 48-57.
37. Gurin, J., *Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth.* SAIS Review of International Affairs, 2014. **34**(1): p. 71-82.
38. Ubaldi, B., *Open Government Data.* 2013.
39. Sheridan, J. and J. Tennison. *Linking UK Government Data*. in *LDOW*. 2010.
40. *Open Government Data*. 2015; Available from: http://www.data.gov/opendatasites.

41. Zuiderwijk, A. and M. Janssen. *The negative effects of open government data-investigating the dark side of open data*. in *Proceedings of the 15th Annual International Conference on Digital Government Research*. 2014. ACM.
42. Chen, C.P. and C.-Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* Information Sciences, 2014. **275**: p. 314-347.
43. PWG, N.B.D., *Draft NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*, in *Reference Architecture*. 2015.
44. Struijs, P., B. Braaksma, and P.J. Daas, *Official statistics and Big Data.* Big Data & Society, 2014. **1**(1): p. 2053951714538417.
45. Sahafizadeh, E. and M.A. Nematbakhsh, *A Survey on Security Issues in Big Data and NoSQL.* 2015.
46. *Dimensions of Big Data*. 2015; Available from: http://www.klarity-analytics.com/392-dimensions-of-big-data.html.
47. Normandeau, K. *Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity*. 2013; Available from: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/.
48. C. Desouza, K. and K. L. Smith, *Big Data for Social Innovation*. 2014: Stanford Social Innovation Review.
49. Hurwitz, J., et al. *How to Ensure the Validity, Veracity, and Volatility of Big Data*. Available from: http://www.dummies.com/how-to/content/how-to-ensure-the-validity-veracity-and-volatility.html.
50. Ali-ud-din Khan, M., M.F. Uddin, and N. Gupta. *Seven V's of Big Data understanding Big Data to extract value*. in *American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the*. 2014. IEEE.
51. Alliance, B.D. *What is Big Data?* 2015; Available from: http://www.bigdata-alliance.org/what-is-big-data/.
52. Vorhies, B. *How Many "V"s in Big Data – The Characteristics that Define Big Data*. 2013; Available from: http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data/.
53. Shan, T. *Big Data Characterized*. 2014; Available from: http://cloudonomic.blogspot.com.es/2014/11/big-data-characterized.html.
54. A. di Paolantonio, J. and B. B. Goewey. *Big Data, What is it Exactly? Datamensional's Take*. 2012; Available from: http://www.datamensional.com/big-data/.
55. Ray, W. *Beyond The Three V's of Big Data – Viscosity and Virality*. 2012; Available from: http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/.
56. Desouza, K., *Realizing the Promise of Big Data*. 2014, Washington, DC: IBM Center for the Business of Government.
57. Krishnan, K., *Data warehousing in the age of big data*. 2013: Newnes.
58. Li, H. and X. Lu. *Challenges and Trends of Big Data Analytics*. in *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference on*. 2014. IEEE.
59. Amodeo, L. *BIG DATA 6V: VOLUME, VARIETY, VELOCITY, VARIABILITY, VERACITY, COMPLEXITY*. 2014; Available from: https://wydata.wordpress.com/2014/12/24/big-data-volume-variety-velocity-variability-veracity-complexity/.
60. Van Rijmenam, M. *Why the 3v's are not sufficient to describe big data*. Available from: https://datafloq.com/read/3vs-sufficient-describe-big-data/166.

61. Wang, R.Y. and D.M. Strong, *Beyond accuracy: What data quality means to data consumers.* Journal of management information systems, 1996: p. 5-33.
62. Basu, M. and T.K. Ho, *Data complexity in pattern recognition*. 2006: Springer Science & Business Media.
63. Farhangfar, A., L. Kurgan, and J. Dy, *Impact of imputation of missing values on classification error for discrete data.* Pattern Recognition, 2008. **41**(12): p. 3692-3705.
64. Wu, X. and X. Zhu, *Mining with noise knowledge: error-aware data mining.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2008. **38**(4): p. 917-932.
65. Chawla, N.V., *Data mining for imbalanced datasets: An overview*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 853-867.
66. Moreno-Torres, J.G., et al., *A unifying view on dataset shift in classification.* Pattern Recognition, 2012. **45**(1): p. 521-530.
67. IBM. *The Four V's of Big Data*. Available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.
68. Zahumenský, I. and J. SHMI, *Guidelines on quality control procedures for data from automatic weather stations.* World Meteorological Organization, Switzerland, 2004.
69. DataONE. *Tutorials on Data Management*. Available from: https://www.dataone.org/sites/all/documents/L05_Exercise.pdf.
70. Arthur, B. *The Difference Between Quality Assurance and Quality Control*. Available from: http://www.dialog.com.au/open-dialog/the-difference-between-quality-assurance-and-quality-control/.
71. *Quality Assurance vs. Quality Control*. Available from: http://www.diffen.com/difference/Quality_Assurance_vs_Quality_Control.
72. Wang, L., G. Wang, and C.A. Alexander, *Big Data and Visualization: Methods, Challenges and Technology Progress.* Digital Technologies, 2015. **1**(1): p. 33-38.
73. Michener, W.K. and M.B. Jones, *Ecoinformatics: supporting ecology as a data-intensive science.* Trends in ecology & evolution, 2012. **27**(2): p. 85-93.
74. Berg-Cross, G.W., Peter and K. Green, *RDA DFT Document 3: Data Processes and Workflow*. 2013.
75. Lenhardt, W.C., et al., *Data management lifecycle and software lifecycle management in the context of conducting science.* Journal of Open Research Software, 2014. **2**(1): p. e15.
76. Fox, P. *Data Management Considerations for the Data Life Cycle*. 2011; Available from: http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_065993.pdf.
77. Emaldi, M., et al., *Linked Open Data as the Fuel for Smarter Cities*, in *Modeling and Processing for Next-Generation Big-Data Technologies*. 2015, Springer. p. 443-472.
78. *Australian National Data Service (ANDS)*. Available from: http://www.ands.org.au/.
79. Kethers, S., et al. *Discovering Australia's research data*. in *Proceedings of the 10th annual joint conference on Digital libraries*. 2010. ACM.
80. Burton, A. and A. Treloar. *Publish My Data: A composition of services from ANDS and ARCS*. in *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on*. 2009. IEEE.
81. Ball, A., *Review of data management lifecycle models.* 2012.
82. Burton, A. and A. Treloar, *Designing for discovery and re-use: the 'ANDS data sharing verbs' approach to service decomposition.* International Journal of Digital Curation, 2009. **4**(3): p. 44-56.
83. *The Bureau of Land Management: Who We Are, What We Do*. Available from: http://www.blm.gov/wo/st/en/info/About_BLM.html.

84. INTERIOR, U.S.D.O.T. *H 1283 1 - DATA ADMINISTRATION AND MANAGEMENT PUBLIC*. Available from: http://www.blm.gov/wo/st/en/prog/planning/planning_overview/guidance/manuals_and_handbooks.html.
85. *CEOS Data Life Cycle Models and Concepts Version 1.0*. 2011.
86. (CSA), C.S.A. *Cloud Security Alliance (CSA)*. Available from: https://cloudsecurityalliance.org/about/.
87. Brunette, G. and R. Mogull, *Security guidance for critical areas of focus in cloud computing v2. 1*. Cloud Security Alliance, 2009: p. 1-76.
88. *DataONE (Data Observation Network for Earth)*. Available from: https://www.dataone.org/about.
89. Michener, W.K., et al., *Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences.* Ecological Informatics, 2012. **11**(0): p. 5-15.
90. Eaker, C., et al., *How information science professionals add value in a scientific research center.* 2013.
91. Tenopir, C., *Research data services: A new focus for librarians.* Consortium on Core Electronic Resources in Taiwan (Concert proceedings 2013), 2013.
92. Higgins, S., *The DCC curation lifecycle model.* International Journal of Digital Curation, 2008. **3**(1): p. 134-140.
93. Constantopoulos, P., et al., *DCC&U: An extended digital curation lifecycle model.* International Journal of Digital Curation, 2009. **4**(1): p. 34-45.
94. Higgins, S., *DCC DIFFUSE standards frameworks: a standards path through the curation lifecycle.* International Journal of Digital Curation, 2009. **4**(2): p. 60-67.
95. *About the DCC*. Available from: http://www.dcc.ac.uk/about-us.
96. *DCC Curation Lifecycle Model*. Available from: http://www.dcc.ac.uk/resources/curation-lifecycle-model.
97. Whyte, A., et al., *Meeting curation challenges in a neuroimaging group.* International Journal of Digital Curation, 2008. **3**(1): p. 171-181.
98. Higgins, S., *Applying the DCC Curation Lifecycle Model*. 2010, cited.
99. Goth, G., *Preserving digital data.* Commun. ACM, 2012. **55**(4): p. 11-13.
100. Ryssevik, J., *The Data Documentation Initiative (DDI) metadata specification.* Ann Arbor, MI: Data Documentation Alliance. Retrieved from http://www. ddialliance. org/sites/default/files/ryssevik_0. pdf, 2001.
101. *Overview of the DDI Version 3.0 Conceptual Model*. 2008; Available from: www.ddialliance.org/system/files/Concept-Model-WD.pdf.
102. *CEOS Data Life Cycle Models amd Concepts Version 1.2*. 2012.
103. DIGITALNZ. *Helping to make New Zealand digital content easy to find, share, and use*. Available from: http://www.digitalnz.org/about.
104. DIGITALNZ. *Make it Digital*. Available from: http://www.digitalnz.org/make-it-digital.
105. Rüegg, J., et al., *Completing the data life cycle: using information management in macrosystems ecology research.* Frontiers in Ecology and the Environment, 2014. **12**(1): p. 24-30.
106. *Report from the Workshop to Improve SDM*. 2011; Available from: http://semanticommunity.info/Other/Scientific_Data_Management_for_Government_Agencies/Report_from_the_Workshop_to_Improve_SDM#Figure_C2._The_generic_science_data_lifecycle.
107. *Stages of the Geospatial Data Lifecycle pursuant to OMB Circular A–16, sections 8(e)(d), 8(e)(f), and 8(e)(g)* 2010.

108. *Federal Geographic Data Committee*. Available from: https://www.fgdc.gov/policyandplanning/a-16.
109. *Creating Knowledge out of Interlinked Data*. Available from: http://lod2.eu/.
110. Auer, S., et al., *Managing the life-cycle of linked data with the LOD2 stack*, in *The Semantic Web–ISWC 2012*. 2012, Springer. p. 1-16.
111. Auer, S., et al., *Introduction to linked data and its lifecycle on the web*, in *Reasoning Web. Semantic Technologies for Intelligent Data Access*. 2013, Springer. p. 1-90.
112. University, M.S. *Records Management*. 2011; Available from: http://archives.msu.edu/records/.
113. *Managing Research Data (JISCMRD)*. Available from: http://webarchive.nationalarchives.gov.uk/20140702233839/http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx.
114. Hodson, S., *Meeting the Research Data Challenge*. 2011.
115. *UK Data Archive*. Available from: http://www.data-archive.ac.uk/about.
116. Faundeen, J.L., et al., *The United States Geological Survey Science Data Lifecycle Model*. 2014, US Geological Survey.
117. Yu, X. and Q. Wen. *A view about cloud data security from data life cycle*. in *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*. 2010. IEEE.
118. Kobielus, J. *Chief data officer: My mixed and nuanced musings on the need for one*. 2014; Available from: http://www.ibmbigdatahub.com/blog/chief-data-officer-my-mixed-and-nuanced-musings-need-one.
119. Ballard, C., et al., *Information Governance Principles and Practices for a Big Data Landscape*. 2014: IBM Redbooks.
120. Schopf, J.M. *Treating data like software: a case for production quality data*. in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. 2012. ACM.