Government (Big) Data Ecosystems: A Systematic Literature Review

Syed Iftikhar Hussain Shah, Vasilis Peristeras, Ioannis Magnisalis

Abstract**—**Data that are high in volume, velocity, veracity, and comes from a variety of sources are usually generated in all sectors, including the government sector. Globally, public administrations are pursuing (big) data as new technology and trying to adopt a data-centric architecture for hosting and sharing data. In the government data ecosystem, adequately executed, big data and data analytics have a direct impact on the way policymakers work. In this research paper, we conduct a systematic literature review. The main aim of this paper is to highlight the essential elements of the government (big) data ecosystem. The essential elements of government (big) data ecosystem include definition, and data types, data lifecycle models. We also discuss gaps in the government data ecosystems literature. We did not find many articles specifically on government (big) data ecosystems and, therefore, borrowed ideas for subject research from various relevant areas like humanitarian data, open government data, scientific research data, industry data in the literature.

Keywords**—**Big data, big data types, big data ecosystem, government big data ecosystems, data-driven government, eGovernment, gaps in data ecosystems, government (big) data, data-driven transformation, public administration, systematic literature review.

I. Introduction

TO become a data-driven organization requires some essential aspects. Such aspects include a *plan* aligned with organization objectives, *people* with appropriate data science knowledge and skills, and *technology*, as well as have a *culture* that understands data and how to use data [1], [2].

Data-driven transformation is a time-consuming and costly task for organizations [3]. However, organizations like Apple, Microsoft, Amazon, and IBM realize the power of data-driven transformation. These organizations are making ample efforts to move their organizations into the data-driven future [1], [2]. They adopted cost-effective, incremental, and sustainable approaches for data-driven transformation to handle big data. They use (big) data and data analytics in sales, marketing, supply chain, manufacturing, and R&D activities [3], [4].

The word "data" derives from the Latin word 'datum' [3]. Data are a discrete, boundless entity that has an unstructured and unprocessed form. Organizations further process such kinds of data to describe relevant objects, ideas, events, or facts [3]. Data are a critical asset in (big) data ecosystem that is understood as symbols and can be either in the form of raw data or in processed data form [5].

Schroeck et al. defined big data as "a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain a competitive advantage in today's digitized marketplace" [6]. Big data is an information asset that is in high volume, velocity, and variety. Big data requires economical, innovative ways of information processing for enhanced insight and decision making [7], [8].

Data are speedily growing from being measured in gigabytes to zettabytes [9]. In the data age, data dominance is the basis of government and industry competition, advantage, and going forward [10]. Data dominance means to acquire more users and customers that lead to more data, whereas more data leads to better products and services delivery of an organization [11].

Government is a sector where the adoption of big data could have a better impact in the relatively short run [12]. The use of big data in public administrations creates various opportunities that include, but are not limited to, drive transformational efficiencies, and offer better public services delivery. Big data is a vital resource for public administration to boost the digital country's economy and renewed job satisfaction for workers [13].

Currently, public organizations are flooded with a vast amount of (big) data generated with high speed [3], through smart data sources such as social networks and the Internet of Things (IoT) [14], [15]. Big data can lead to the new data-driven government for the public administration, new knowledge for informed decision making for the decision-makers in government, introduce new products, processes, markets, and to drive value creation, i.e., data value creation for the businesses [12].

Big data also provides an efficient way of feedback to public administration for necessary amendments to their policies as per the current needs of stakeholders [16]. Data provide a useful way of feedback to government organizations for essential modifications in their policies as per the current needs of stakeholders [16]. In the literature, we have found different types of big data. The types of data include machine-generated data, human-generated data, web and social data, biometric data, and transaction data [14]. A detailed overview of the types of big data is described in the forthcoming section of this research article.

Big data is generated from numerous sources. Such sources

Syed Iftikhar Hussain Shah is Researcher and Ph.D. Candidate in eGovernment/Information Systems Management is with the School of Science & Technology, International Hellenic University, 14th km Thessaloniki - Moudania, 57001 Thermi, Greece (phone: +44-7932606032; e-mail: i.shah@ihu.edu.gr).

Dr. Vasilis Peristeras, Assist. Professor, is with the School of Science & Technology, International Hellenic University, 14th km Thessaloniki - Moudania, 57001 Thermi, Greece (e-mail: v.peristeras@ihu.edu.gr).

Dr. Ioannis Magnisalis, Associate of Assist Professor, is with the School of Science & Technology, International Hellenic University, 14th km Thessaloniki - Moudania, 57001 Thermi, Greece (e-mail: i.magnisalis@ihu.edu.gr).

of big data may be either internal or external to the organization [3], big data may reside in the structured, semi-structured or unstructured form [3], [17], [14], it may be open to all or in use within a group of organizations [18], [19]. Data may be stored centrally in the form of registers [20], and generated in real-time or batch mode [14].

Currently, online social networks like Twitter, Facebook, LinkedIn, Instagram, and the IoT are creating a massive amount of data. In the literature, we noted about 2.77 billion social network users and 26.66 billion IoTs connected devices installed worldwide [21]. Social media users are regularly exchanging information through social networking platforms, whereas entities across the globe are using IoT devices to collect and exchange required data about various sectors, like health and agriculture, from the desired environments [22]. The data ecosystem consists of different elements, including people, process, technology, organizations, infrastructure, and data lifecycle. In the literature, the data ecosystem is defined as a complex of interrelated components that deal with the evolving data, models, and supporting infrastructure during the whole data life cycle [23]. The data ecosystem is established based on collaborations among stakeholders. The stakeholders include public and private organizations, development partners and users.

The importance of data is growing as the fuel of the new data economy, and the role of the data ecosystems is also growing at the same pace [23].

We focused on three essential elements of the government (big) data ecosystem. These elements include the definition of (big) data ecosystems, types of (big) data used in the government data ecosystem, and (big) data lifecycle models. A detailed overview of the aforesaid essential elements is described in the forthcoming Results section of this article.

In the literature, we noted that (big) data is progressively getting the attention of professional multi-national international organizations, digital ministries, and governmental departments across the globe. We found that the (big) data ecosystem is one of the new hot research topics within the research community.

We identify gaps in the literature regarding critical elements of (big) data ecosystems. These literature gaps include no well-established definition of (big) data ecosystems, no extensive work on the classification of types of government (big) data, and no data lifecycle for the next generation data-driven government. A detailed overview of the gaps above is described in the forthcoming section of this research article.

In the research literature, we found articles on various areas based (big) data ecosystems. The published articles are based on areas like scientific research, Semantic Web & web content management, open government, business, etc. We did not find many research articles specifically on the government (big) data ecosystem. Therefore, we borrowed ideas for the government (big) data ecosystem from the existing literature on the areas above based (big) data ecosystem.

The remainder of the research paper is organized as follows. In Section II, we present the research method and search process and its outcomes. In Section III, we showcase the results of the literature review. In Section IV, we include a detailed discussion on the findings, and lastly, in Section V, we present our conclusion.

II. Research Method

In this section, we present research goals, research question, and explain the search process to find relevant research publications to conduct a systematic literature review on the subject area of research.

A. Goal and Research Question

The goal of our research is to highlight the essential elements of the government (big) data ecosystem. For this, we framed a key research question – RQ: What are the essential elements of government (big) data ecosystem? The question aims to explain the essential elements of the government (big) data ecosystem. Moreover, in the discussion section, we also discuss gaps in the government data ecosystems literature. We attain our research goals and find out the answer to the above question by conducting a systematic literature review, with the results detailed in the results section.

B. Search Process

In this research article, we adopted a systematic literature review as per the guidelines proposed by [24]. To cover a broad spectrum of relevant publications, we searched for relevant articles from the following widely recognized and extensively used electronic libraries. Electronic libraries include ACM Digital Library, IEEE Xplore Digital Library, Science Direct, and Springer Link.

Our search process consists of the following four steps. These steps were not scheduled in advance and were taken during the search process to cover the relevant topics of government (big) data ecosystems as per the goal of this research article. Firstly, we performed searches on different main keywords strings like "GOVERNMENT DATA ECOSYSTEM", "DATA-DRIVEN GOVERNMENT," "DATA ECOSYSTEM" and "DATA-DRIVEN PUBLIC ADMINISTRATION" with options "exact phase" and "matches all." However, we found a limited number of research articles in some of the aforesaid electronic libraries. Secondly, we performed searches on the keyword strings above with options "matches any" and "with at least one of the words." In the second attempt, we found more research articles as compared to the first search attempt. Nevertheless, we observed that search results from the previous two search attempts did not cover critical associated terms/components regarding government (big) data ecosystem. Thirdly, we performed an additional search on keywords strings like "data-driven policy," "master data management," "data portals," "base registries," "modern data architecture," "data analytics. Lastly, we performed the following: i) explored some advanced countries, including EU Member States official websites, to find out (big) data policies & strategies documents; ii) searched for international consultancy firms like Gartner, IBM, and KPMG relevant case studies reports. And, we also studied research articles that were already referred to in the research articles found in the search above process. We called these research

articles as "secondary references".

We did searches on keyword strings in the research paper title, abstract, and keywords. We confined our search to publications written in the English language and gave top priority to select content type 'only journals'. However, research articles from other content types like international conferences, government (big) data policy documents, and international consultancy firms were included to have enough relevant research publications. The procedure for selecting the literature was conducted in February 2019 and in October 2019.

The final number of papers that were gathered after removing duplicates is 909. Then, papers having titles irrelevant to the research were also identified manually and excluded in the next review phase; thus, it reduced the papers to 499. The above-mentioned procedure was repeated by scanning the papers' abstracts, reaching 347 papers. The authors of this study read the full text of these papers, culminating in 285, to extract information and answer the proposed research questions. The results of the search process are shown in Fig. 1.

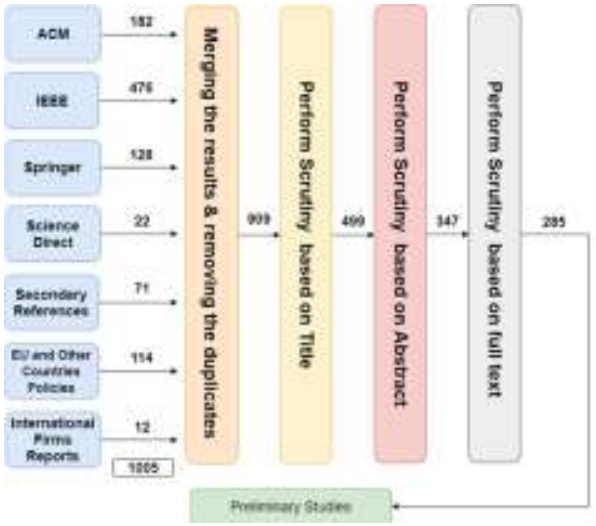

Fig. 1 Procedure for identifying preliminary studies

III. Results

In this section, we describe the outcomes of the literature review regarding the government (big) data ecosystem. The results are organized as per the proposed research question, and each element, mentioned in Table I, are examined thoroughly and separately. We also mention the total numbers of papers for each element that are included in the preliminary studies.

TABLE I
Research Question: Essential Elements of the Government (Big) Data Ecosystem

Elements	Definition of a (big) data ecosystem	Types of Data	Data Lifecycle models
Total papers	32	20	64

A. Results for RQ - Key Essential Elements of Government (Big) Data Ecosystems

As per the literature, the authors focused on three essential elements of the government (big) data ecosystem. These elements include the definition of (big) data ecosystems, types of (big) data used in the government data ecosystem, and (big) data lifecycle models. The detail of the essential elements of the government (big) data ecosystem is as below:

1. Definition of (Big) Data Ecosystems

We have found about 25 research studies, which have attempted to define the data ecosystems. Here, we present some definitions of data ecosystems found for the data ecosystem concept in different contexts. References [25] and [26] define the data ecosystem as a network of humanitarian actors, governments, and private sector organizations, and affected communities in which they interact with each other to produce, collect and analyze digital data about vulnerable populations. Humanitarian data ecosystem actors exchange data for disaster management through close coordination. Reference [26] added that sources of crisis-affected communities data include mobile phone records, social media posts, satellite imagery, sensor data, financial transactions. The said definition pertains to the humanitarian domain. Parsons et al. define the data ecosystem, specific to the scientific research field, as "the people and technologies collecting, handling, and using the data and the interactions between them" [27]. References [28] and [29] define data ecosystems as socio-technical complex networks in which actors (organizations and individuals) interact and collaborate to exchange and use data as the primary source to foster innovation, create value, and support new businesses. Reference [31] defines the (big) data ecosystem as a heterogeneous network of software, hardware, and networking resources, human capital (such as skills), industry applications and methodological techniques, social actors, and the new ideas and concepts those actors coin.

In the literature, data ecosystem definitions are more explicit to the open data ecosystems [30], [31], [39]. The open data ecosystem is a holistic set of activities for releasing and publishing open data on the internet, where open data users can conduct various activities [30]. The examples of data activities include searching, finding, evaluating, cleansing, analyzing, enriching, combining, linking, visualizing, interpreting, and providing feedback. Dawes et al. define open data government ecosystems as a way to assess existing conditions and figure out what changes in policies, strategies, practices, relationships, or participants of relevant stakeholders (government, private sector, community) would bring a government's open government data program closer to producing desired benefits [32]. Smith et al. define the open data ecosystem as a loosely coupled network of different open data actors that interact with each other and perform open data processes such as create, open, find, analyze, process and discuss and provide feedback [33].

In the literature, we also observed that data ecosystem definitions pertain to different specific domains. The examples of such domain-based data ecosystems include the humanitarian data ecosystem [26], [25], [34], scientific research data ecosystem [27], personal data ecosystem [35]. Reference [36] stated that these definitions referred to an environment domain where an ecosystem emerges. These domains may also

determine the associated organizational structure and its rules. Moreover, authors of relevant research articles named the domains and performed their research work in the respective domains.

2. Types of (Big) Data

To describe the literature on types of (big) data in a better presentable way, we organize the types of (big) data in six classifications, as shown in Fig. 2.

Fig. 2 Classification of Types of (Big) Data

The explanation of each classification and its relevant types of (big) data is as below:

Structured, Un-structured and Semi-structured

This classification of structured, unstructured, and semi-structured data types is helpful to determine how the incoming data needs to be processed.

The *structured data* possess highly organized schema, standard format, and layout [3]. The structure data are stored in databases [15], accessed, and processed in an organized and precise manner [3]. Structured data are stored, accessed, and processed in a specific format and an ordered manner [15]. Examples include library catalogs, census records, phone directories, and databases. Structured data account for about 20% of the total existing data. Such type of data are used the most in computer programming for data collection, data cleansing, data analysis, and data visualization software tools [15].

Un-structured data are not abided by predefined format, structure, and data model. Therefore, it is difficult to convert or map unstructured data into the format required for efficient processing [3]. Examples include X-Ray and satellite images, audio files, videos.

Demchenko and Membrey were of the view that scientific research data also include un-structured data. Such data are collected from scientists' observations or experiments. Scientists call it raw data, and they further process and use this data for their scientific research works [37].

Unstructured data have no clear format [15] in storage. To further process structured data, data experts need to pass it through different phases of the data lifecycle, which is a time-consuming and tedious job [15]. Data experts process big data by using appropriate data analytical tools or process data manually. There is always ambiguity between unstructured and semi-structured data as mostly semi-structured data appear to be unstructured.

Semi-structured data are not in the same format as structured data but contain some elements like semantic tags that make it easier to analyze. An example of semi-structured includes HTML code that does not limit the amount of information to be collected in a document. However, HTML code still enforces a certain hierarchy through semantic elements [15], [38]. Other examples of semi-structured data include XML, JSON documents, NoSQL databases, Electronic data interchange (EDI), and RDF.

In the literature, we also noted that unstructured and semi-structured data make up about 80% of the total existing data. Semi-structured data may be irregular or incomplete and have a structure that may change rapidly or unpredictably [3].

Real-Time Data and Batch Data

It is important to know whether the data are analyzed in real-time or batched for further analysis [14].

Real-time (big) data are generated in real-time by systems [14]. Examples of real-time data include stock market data, and sensors or Twitter feed data, traffic data, radar data, bank ATM data, and weather data [39], [40]. Presently, there is a data management system that is based on real-time data and known as a real-time data management system. Such systems are providing suggestions to individuals, businesses, and governments on different matters. People take decisions-based on suggestions by the real-time data management systems. For example, [39] introduced a traffic navigation system. Mobile vehicles send time-constrained requests to the navigation system. The navigation system processes the said request based on real-time traffic data and transmits the best possible routes to the requesters [39]. The real-time data help the people to reach the destination through vehicles, particularly when they have specified deadlines on their arrival time [40].

Batch data are assembled with some interval of time. In batch data systems, the interval of time may range from a few minutes to hours [39], [41]. Examples of batch data include payroll data, image processing, and billing data. Batch processing is an efficient way of processing (big) batch data. In batch data processing, data are loaded in huge batches at a specific interval. The users use job batch data until the next job data load occurs during batch job processing [14]. Batch data processing is extensively used in the production of high-value, typically low-volume materials, including pharmaceuticals and microelectronics. The economic cost of lost process performance is typically high and has motivated extensive research in batch process monitoring, fault detection, and control. However, Wang et al. stated that data-driven methods have an essential role in this area [41]. High-availability distributed object oriented-Hadoop, an open-source software framework, is focused on batch data processing [42], [43].

Human-Sourced Data, Process-Oriented Data, and Machine-Generated Data

Such classification of (big) data type is helpful to determine the scope from a business perspective [10].

Human-sourced data are a record of human experiences, previously recorded in books and works of art, and later in

photographs, audio, and video [14]. Human-sourced information is now almost entirely digitized, using modest tools and technologies, and stored everywhere from personal computers to social networks. Examples include social networks (Facebook, Twitter), Internet searches, and consumption data of water, gas, and electricity generated by the smart meters.

Process-oriented data: This type of data is the outcome of traditional business systems. These processes record and monitor business events of interest that include registration of a customer, manufacture a product, and take an order [15]. The data collected through process-oriented data are highly structured in nature and usually stored in relational database systems. Examples of process-oriented data include medical records, data produced by businesses, banking/stock records, and data produced by public agencies [15], [14].

The *machine-generated data* are produced from IoT devices. Machine-generated data have a size and speed that is beyond traditional approaches [14], [44], and mostly derive from sensors and computer systems. The sensor-based data sources include weather sensors, traffic sensors, and security/ surveillance cameras whereas computer systems data sources consist of computer logs and weblogs [15]. In the literature, we observed that a huge increase in the number of sensors in the world results in growing data volumes and it is considered as a crucial factor of the information that is warehoused, and processed by the data actors including businesses [10], [37], [44]

We also noted ownership of human-sourced data, process-oriented data, and machine-generated data as a critical challenge. This challenge requires the special attention of the researchers to conduct further research in the future.

Government, Business and Society Data Including Personal Data

This classification consists of (big) data types that are managed by public sector organizations. However, such data relate to citizens and businesses, as well.

Government data: Digital data that a government ha have in its all public organizations [20]. Government data can either be generated or collected by the public administration. Government data are one of the critical assets of the State [20]. The examples of government data include residents' social security numbers, public health data, criminal data, public scientific data, electoral rolls, and vehicle registration data. Government data include shared data that are available to the rest of public entities. Such government entities use this shared data to achieve greater efficiency in their organizations. Government data also have a dataset that is shared within a smaller set of public entities instead of the whole government. This set of public entities belong to a specific business vertical or have a typical goal/project [20], [35].

Business data are the information that is used to plan and operate a business organization. The business data contain information about customers, places, products, business rules, and market trends. An organization usually extracts patterns or forecasts customer behavior from big data [45]. The examples of business data include customer and product data, transactional data, and marketing data. Technological solutions help organizations to gather, store, and track business data. For example, in the past, company salespeople used a Rolodex (business card holder) to store customer contacts. However, in the current age, sales department staff uses state-of-the-art CRM solutions to store customer contacts.

Commercial entities are capturing business (big) data from different emerging sources, including social networks. The examples of such business data include consumer's needs, consumer behaviors, and customer satisfaction about brand products [28], [46], [47]. The volume of business data is increasing at a very high speed. For example, Amazon is processing more than 35 transactions every second [28], [47].

Society data: Consists of data about vulnerable societies that are collected, produced, integrated, analyzed, shared, and published by the humanitarian actors, their partners, and affected societies [26]. The examples of society data include flood prediction data, earthquake-prone areas data, poverty data, illiteracy data, and disease virus data, and human trafficking data. The societies are using society data to enhance their ability to address societal and cultural problems [26], [25], [34].

Moiso and Minerva define *personal data* as data about individuals, their behavior, and their actions [35]. In 2009, Kuneva, the former European Consumer Commissioner, stated in her speech that personal data are the new oil of the internet world [48]. Examples of personal data include personal photos, videos, chat, credit card numbers, addresses. Public administrations maintain such data about citizens and business entities [20]. Personal devices, like smartphones and tablets, generate personal data. These data are assumed to become the "energy" or the "new currency" for the digital world. Personal data about people and their activities also offer new opportunities to people and organizations. In the literature, we observed the following two significant challenges with personal data. The first challenge is inadequate personal data privacy by internet service providers, while the second challenge is a lack of users' control on their personal uploaded data [35], [49], [48].

Open and Closed Data

Open data are a piece of data that is freely available to any individual for use, reuse, and redistribution subject only, at most, to the requirement to attribute and share-alike [18], [19], [50]. The advantages of open data have been widely recognized around the world, and there has been a tendency in many countries to open the data of the public sector. The government is an open data provider; however, private companies could also open a part of their data [51]. Open data offers new business openings for the business entities and related actors [51]. European open data portal contains different open datasets that include national public procurement notices, national retail trade-monthly, national economic and financial data, national education, culture, and sports data, energy data, health, public scientific research data, and medical data [52].

In the government sector, open data are expected to bring many advantages, such as stimulating citizen participation and innovation, avoiding data duplication, allowing competent

decision making, stimulating transparency, and stimulating economic growth [28] whereas in the private sector, the opening of data can provide other benefits, such as new partners, new interests in the company's primary products/ services, new kinds of business activities, and new customers for the product/service [51].

Closed data are the type that government organizations use within their premises and do not share with third parties [3]. Closed data in businesses include revenue data and product formula data, whereas closed data in the government sector include employee service records, employee performance assessment reports [53], [3]. Closed data cannot be shared with any third party, and its access is restricted to a data owner or a particular group in a public organization due to security constraints and relevant policies [50]. Public organizations need security and privacy measures for closed data secrecy and confidentiality [3]. Examples of such measures include the implementation of intrusion detection systems, firewalls, and enforcing access control policies [52].

Master, Reference, Metadata, and Transactional Data

Master data are a single source of primary business data used across multiple systems, applications, and processes [54]. According to Gartner, master data consist of a uniform and consistent group of identifiers and attributes that illustrate the main entities of organizations like customers, products, and suppliers [53].

Master data represent the critical transversal entities of the enterprises, particularly the business, which provides a context to the transactions [20], [77]. The examples of master data include *citizen master data* (a citizen's social security number, name, and address, etc.), *product master data* (product id#, product name, product unit), *customer master data* (customer id#, customer name, and address, etc.) [55], [56]. The master data make up the fundamental data entities of the government to achieve a data-driven administration [57], [20].

Reference data are the set of permissible values to be used by other data fields [20]. Zornes defined reference data as coded, semantically stable, relatively static data sets shared by multiple constituencies (people, systems, and other master data domains) [58]. Data users re-use reference data, and it gains value.

Typically, reference data do not change overly much in terms of definition (apart from occasional revisions) [18], [17], [20]. Errors in reference data affect the quality of master data and relevant transactional systems as well [58].

The examples of reference data include ISO country codes, ITU country phone codes, and internet country code top-level domains [20], [58].

Metadata is data that illustrates other data. Metadata is a piece of structured information that describes, explains, locates, makes it easier to retrieve, use, and manage information resources [59]. It is an integral element of data management, as it facilitates the discovery and efficient use of the data. Metadata should be created along with the dataset in a standardized way, published alongside the data, and updated whenever the dataset is updated [18], [17]. For example, book metadata includes the author's name, book title, publisher, edition, date of publication, number of pages, and much more. There are different kinds of metadata. The first kind of metadata is descriptive metadata. Descriptive metadata contains elements that support to discover and locate information resources. Descriptive metadata elements include title, abstract, author, and keywords. The structural metadata expresses information about containers of data and implies how compound objects are assembled. The structural metadata describes types, versions, relationships, and other characteristics of digital materials.

Administrative metadata provides information that helps manage a resource. Examples of such administrative metadata information include when and how metadata is created, file type, and other technical information, as well as who can access it [60].

Transactional data are obtained from business transactions. For example, when a firm sells a product or a customer purchases a product from an organization [61]. Transaction data are vital for business analytics and business intelligence. The examples of transactional data include billing data and sales data. The transactional data contain some essential attributes, like transaction ID, and transaction date and time, to uniquely identify the business transactions. Transactional data are more volatile than master data due to the frequent creation and changes in it. Generally, master data do not need to be changed or created with every transaction [61].

There are a few studies on types of big data [14], [50]. Each research study considered some specific types of (big) data on their own. Moreover, in the literature, authors only proposed classifications of types of (big) data on their selected types of big data. Most of the (big) data research studies published articles in blogs, online articles, and commercial organizations websites. We did not find many research articles in well-known journals, specifically on the classification of types of big data. In the above paragraphs, we tried to describe existing literature as extensive work on the classification of types of big data as well.

3. Data Lifecycle Models

The Data Lifecycle Model and Its Importance

The (big) data lifecycle represents all phases of (big) data during its life, from its creation to its distribution, use, and reuse [62]. The data life cycle provides a high-level overview of the stages involved in the successful management of (big) data for use [63], [13]. In the literature, we found and read 64 research articles presented in some way different data lifecycle models. In these articles, we identified 46 different data lifecycle models presented during the last 25-year period. The identification of the data lifecycle that fits the company's data usage is a critical task for the organizations. However, the method of data lifecycle management is vital to transform data into knowledge [64] and to extract the requisite value from data big data to improve organization data operations [65].

The documentation and proper management of the data lifecycle can help to identify dataflows and work processes for stakeholders in the data ecosystem. The data lifecycle supports public organizations to tackle the problem of data management

[13].

Existing Data Lifecycle Models and Their Categories

As explained earlier, in the literature, we found 46 different data lifecycle models, which were divided into the following three main categories. This categorization of the data lifecycle is helpful for the readers of the papers, the research community, and us as well, to apply these literature models in future research work.

i. In the first category of data lifecycle models, researchers just described the phases as terms, and no formal research is carried out by them. For example, [35] described the personal data lifecycle model phases in a scattered way. We did not find formal research work on such models in the respective research studies. The authors did not consider a unique phase in their proposed data lifecycle models. The examples of other such data models include the eLearning data lifecycle model [66], data common lifecycle [67], data lifecycle model [53], [32], [28], open data model [68], [69], digital libraries lifecycle model [70], and knowledge management model [71].
ii. In the second category of data lifecycle models, researchers adopted a formal approach to introduce their models. These models provided a basis to propose a model for the next generation data-driven government. For example, [72] presented described data collection, data publication, and data re-use phases, whereas the model of [73] mentioned data planning, data access, data archive, and feedback phases. We also describe the explanations of each model in the subsequent section. The examples of other such data models include Web Content Management [74], IBM data lifecycle model [75], DataONE Data Lifecycle [63], CIGREF data life cycle [65], Research Data Lifecycle (UK) [76], USGS [77], Hindawi Data Lifecycle model [78], El Brass Big Data lifecycle models [65], [79]. We also consider these models in our future research work, as highlighted in the forthcoming conclusion section of this paper.
iii. In the third category of data lifecycle models, researchers introduced a few unique phases in their data lifecycle models. For example, the PII data lifecycle focused on the data security phase [80], whereas the data lifecycle model of [81] focused on the data governance phase. The examples of other such data models include the General Data DCC Curation lifecycle model [82], DDI version 3.0 Combined Life Cycle Model [62], ANDS Data Sharing Verbs [83], Capability Maturity Model for Scientific Data Management [84].

Explanation of Some Data Lifecycle Models

We explain a few data lifecycle models to provide more detailed understanding. For example, *IBM data lifecycle model* phases include data creation, data use, data sharing, data update, archive, store/retain/dispose of with three layers of data management. IBM introduced an additional three layers in its data lifecycle model. These three layers of the IBM model include test data management, data masking, and archiving [75]. In the test data management, technicians automate the creation of realistic, rightsized data sources that mirror the behaviors of existing production databases. Data technicians perform this action during the process of developing new data sources. Data technicians also create a subset of actual production data and reproduce actual conditions to find problems in the testing cycle at an early stage. In the data masking layers, organizations mask sensitive information like the personal information of an individual to protect data privacy. In the data archiving layer, data are archived based on specific parameters or business rules like the age of the data. It helps storage administrators to formulate an appropriate storage strategy to archive inactive data in their respective data repositories.

The scientific research field-related organizations adopt the *DataONE lifecycle model*. The DataONE lifecycle model phases include plan, collect, assure (format, coding, quality assurance, quality control), describe, preserve, discover, integrate (transform), and analyze. The DataONE life cycle supports the full data lifecycle and enables new and data-intensive disciplines, i.e., ecology discipline, to manage the data [63].

USA (NIST) – [big] data lifecycle model phases include collection, preparation, analysis, and action (analytics, visualization, access) [85]. Reference [86] introduced this data lifecycle for (big) data. This model is more focused on the data analytics part as compared to data planning, data filtering, and data enrichment. The said lifecycle did not cover the planning, quality, security, archive, and data governance phases.

Media production discipline-related organizations implement the *media production lifecycle model.* This model consists of phases include premeditate, capture, archive, annotate, query, message construction, organize, publish, distribute [86]. This model contains basics phases that usually exist in most data lifecycle models. Moreover, this model did not cover some crucial phases like planning, storage, quality, feedback, and governance phases.

Knowledge lifecycle for e-learning is a semantically driven knowledge management lifecycle. These model phases include knowledge acquisition, knowledge modeling, knowledge reuse, and knowledge maintenance. This model helps to manage semantics and knowledge in the leaning domain and demonstrates its value through knowledge re-use. This model lifecycle is focused on knowledge management and did not describe other phases like phases of data above lifecycle models.

Phases of Data Lifecycle Models

In 46 different data lifecycle models, we found 240 phases. This total number of phases also contains some phases that exist in more than one data lifecycle model. The examples of phases of data lifecycle models include planning [79], [87], [65], study concept [62], collection [72], [77], [14] creation [73], capture [86], receive [65], integration, filtering [79], classification [81], process [88], [20], enrichment [37], [84], [89] visualization [90], [85], storage [82], [83], [76], access [88], [87], use [30], [35], re-use [72], [88], feedback [73], archive [37], destruction

[91], quality [92], [81], data security & protection [88], and data governance [81], [78].

Explanation of Data Lifecycle Model Phases

Here, we explained typical phases of 46 data lifecycles models as an example for the familiarity of the readers of this paper:

Hardman defined the *planning phase* as two different processes, and the first process is called premeditate, and second is known as message construction. The premedicate means planning of creating a single unit of data, whereas message construction explicitly means to plan and intent to combine several pieces of data to a larger whole [86]. In the literature, we found some data lifecycles that contain 'planning' as a first phase [77]. Michener et al. stated that the planning phase could be helpful for researchers, including scientific researchers, to save time, enhance research efficiency, and to fulfill the requirement of research sponsors regarding planning for data management [13], [76].

The *Data Collection* phase consists of a set of activities through which data is collected from different internal and external sources and in various forms [77] whereas the *Data integration* phase consolidates these data silos into a single place/platform [93], users can do queries easily and obtain a response from this single data source. References [17], [93], and [94] called data integration as a database subarea that provides uniform access to various data sources. Reference [65] stated that this phase is based on a set of rules and policies. In the literature, we found various similar terms of the integration phase. Examples of such terms include ingestion, aggregation, curation, and repurposing [65], [17].

The *Data filtering* phase focuses on the purification of data by filtering the noisy and erroneous data [78], [65]. This phase also allows data classification in different formats [78]. The filtered data are further processed through succeeding phases. After due process, policymakers use filtered data to make better decisions within a limited time and with fewer resources [81], [88].

The *Data enrichment* phase focused more on data valuation [37]. Public organizations can achieve data values by incorporating additional information in the collected data. The enrichment phase has a positive impact on the results of the analysis phase.

Data analysis is the most common phase of all data lifecycle models. Khan et al. stated that the analysis phase enables an organization to handle ample information that can affect the business [78]. Jennifer et al. and Kaufmann were of the view that the analysis phase is like a human brain that processes the information for the next appropriate required action by human beings [89], [93]. The visualization phase deals with presentation and visualization of the results, as well as interpretation of the meaning of the discovered information from the big data [90], [85], [95].

In the *Data destruction* phase, duplicated data, unnecessary data, and useless data are removed from the system [91], [88]. We may consider data in terms of the end of life [88].

The *Data quality* phase helps to maintain data quality during data collection, data integration, data analysis, data publishing, and data sharing [90].

Data privacy, security & protection phase focuses on data security & protection in terms of data integrity, access control, and privacy [65]. This phase exists in all phases of the data lifecycle, i.e., from planning to archive, to maintain data security & protection [88].

The *Data governance* phase consists of overall data action supervision throughout the data lifecycle model [88]. Data governance does not only include consistent management, but it also includes relevant holistic policies, guidance, processes, and overall decision making. The said phase also exists in all phases of the (big) data lifecycle models. [81], [88].

In the literature, we observed that some data lifecycle studies focused on specific (big) data aspects. Such data aspects include data quality [65], [77], data processing [17], data conception [82], data management [77], data security [91], data strategy [88], [20], analytic [85], [95], feedback and data refinement [87], [96]. So far, we could not find a specific data lifecycle model for government (big) data ecosystems. We also propose a data lifecycle model for government (big) data ecosystems as one of the topics of future research.

IV. Discussion

In the literature, we observed that (big) data is progressively getting the attention of the academic researchers, professional multi-national international organizations, digital ministries, and governmental departments across the globe. We also highlighted gaps in the literature about crucial elements of government (big) data ecosystems. The literature gaps provide avenues for government (big) data ecosystem researchers and public administrations to conduct further investigations. We also provide detail on the topics above in this section as below:

In recent years, universities also introduced degree programs that are majorly focused on big data. Becker highlights that "big data is the oil of the 21st century" as the ability to refine and use big data has become a critical success factor for government, businesses, and communities [97].

The (big) data ecosystems have a vital role in smart cities. Shah et al. stated that globally, urban cities' populations are expected to be increased by about 68% by 2050. Such an increase in the human population will create a necessity for better services and suitable infrastructure for people of the urban cities [98].

The concept of smart cities is one of the possible solutions to tackle the problems that happen from population growth and other critical social and environmental challenges. Other such challenges include environmental pollution, and shortage of energy sources [98]. IoT, big data, and appropriate communication infrastructure for electronic devices are the core components of smart cities. In such cities, a vast amount of (big) data would be generated through intercommunication among billions of electronic devices. Moreover, implementation of the (big) data ecosystem would be one of the key aspects and a game-changer towards a state-of-the-art and disaster-resilient smart city [98].

Machine learning, artificial intelligence (AI), IoT natural

language, and 'bots' (automated scripts and programs) are becoming standard elements of big data and technology stack strategies. The efficiencies of the aforesaid emerging technologies combined with the underlying power of big data analytics are impacting all sectors, including public administration, in a short period [99]**.** Data (big) are a foundational driver of digital transformation, as well as an enabler [100].

A. Gaps in the Literature about Key Elements of Data Ecosystems: In the literature, we identified the following data ecosystems gaps:

1. No Well-Established Definition of Data Ecosystems

There is no well-established definition of data ecosystems [20], [36], [101], including government data ecosystems. We have found about 25 research studies, which had attempted to define the data ecosystem. However, the authors of the papers use the term data ecosystem without providing a detailed concept [81], [101]-[103]. Some authors [32], [28] referred to generic ecosystems' definitions in their work to fill the definition space and to present their core topic of research. For example, they referred to ecosystems [31], [103]-[105] business ecosystem [106], [107], innovation ecosystems [108], information system ecosystem [109], software ecosystem [110], and digital ecosystem [111].

2. No Extensive Work on the Classification of Types of Government (Big) Data

In the literature, researchers mainly explored technical aspects of the data ecosystem that reflect a small portion of the data ecosystem [37], [29]. We have similar observations on existing literature about data ecosystems, including the classification of types of big data. There are a few studies [14], [15] on 'types of big data', and each research study considered some specific types of data on their own. Moreover, in these research studies, they proposed classifications on their selected types of big data. We found a few research studies, like [59], on 'types of big data' in blogs, online articles, and web pages of some commercial organizations websites. We did not find a specific research article in well-known journals on the classification of types of big data. There is a lack of extensive work on the classification of types of big data, as well.

3. No Data Lifecycle for the Next Generation Data-Driven Government

In the literature, we found specific area-based data lifecycles. For example, [72], [83], [76], [77] described data lifecycles based on area 'scientific research' whereas, [86] mentioned the data lifecycle based on the area of 'media production'. In the literature, other data lifecycles are based on areas such as Semantic Web [73], databases [80], open data [17], information systems [71], and cloud computing [91]. In the research literature, some data lifecycle studies focused on specific (big) data aspects such as data quality [65], [77], data processing [17], data conception [82], data management [77], security [91], data strategy [88], [20], analytic [85], and feedback and data refinement [87], [96]. As far as we know, there is no specific data lifecycle on data-driven government.

4. Confusing Literature on the Data Actors and Their Roles

During the literature review, we observed various redundant types of actors in data ecosystems. The types of actors and their roles in the data ecosystem differ from one research work to another. The notable variances of actors and their roles in the data ecosystem are due to assigning different actors titles and their contradictory roles described in the various research articles [28], [35], [112], [113]. In the literature, it is challenging to find a set of actors and their roles in the data ecosystem. There is no single research article to explain the dimensions of actors and their roles in the government data ecosystem.

5. No Specific Research Articles on Government (Big) Data Ecosystems

There are a limited number of academic papers on data ecosystems in general, and we did not find research papers specifically on government data ecosystems.

The above discussion regarding data ecosystem gaps in the literature is the questions/problems and raises the attention of data science researchers to explore the respective areas further so that they may address or answer these questions adequately.

V. Conclusion and Future Work

In the (big) data-driven world, data dominance is the basis of governmental and commercial organizations' competition, efficiency, and organizations' better future. Government (big) data ecosystems can help the public administration to make evidence-based decisions, ensure data interoperability and data privacy, prioritize the problems, encourage civic participation in government in developing holistic policy development processes, and contribute to a better government.

State-of-the-art research in government (big) data ecosystem is still far behind in maturity and has an enormous scope of research; continuous and active research participation is required at our end to have new findings.

The key motivation of this research paper is to contribute and present some findings on government (big) data ecosystems. The authors conducted a systemic literature review on essential elements of the government (big) data ecosystem instead of focusing its technological elements.

During this research work, we have found out about 25 research studies that have attempted to define data ecosystems. We have studied 46 data lifecycles models presented during the last 25-year period, and we also found numerous types of (big) data. The overview regarding data lifecycle models will benefit public organizations to better understand and choose or devise an appropriate data lifecycle as per their visions and needs about (big) data management.

We also highlighted critical gaps in the literature about government (big) data ecosystems. The gaps include no well-established definition of the (big) data ecosystem such as government data ecosystems, no data lifecycle model for the next generation data-driven government, the challenge to find and define a set of actors and their roles, and as well, we did not

find a single paper specific to government (big) data ecosystems.

This paper highlighted the fundamental elements of government (big) data that need to be studied for further progress in the research of this subject.

As future work, we intend to explore further and analyze the literature to address the aforesaid identified gaps on the subject ecosystems by proposing: i) definition of a (big) data ecosystem including government (big) ecosystem, ii) a framework for the actors and their roles in the government data ecosystem, and iii) a data lifecycle for the next generation data-driven government. We also intend: iv) to develop a Critical Success Factor (CSF) framework for the government data ecosystem incorporating six leading data ecosystem CSF dimensions: Organizational, Economic, Social, Technical, Legal, and Semantic (OESTLS), v) clearly defined a set of components of the government data ecosystem based on an adequate criteria, and lastly, vi) a theoretical framework of data-centric (conceptual) architecture for the next generation data-driven government.

ACKNOWLEDGMENT

The European Union-funded project Co-Inform, Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) H2020-SC6CO-CREATION-2016-2017 (CO-CREATION FOR GROWTH AND INCLUSION), Grant Agreement 770302 and The European Union-funded Project: Digital Europe for All (DE4A), Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020), H2020-SC6-GOVERNANCE-2018-2019-2020 (GOVERNANCE FOR THE FUTURE), Grant Agreement: 870635.

REFERENCES

[1] A. Fabijan, P. Dmitriev, H. H. Olsson, and J. Bosch, "The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale," *Proc. - 2017 IEEE/ACM 39th Int. Conf. Softw. Eng. ICSE 2017*, pp. 770–780, 2017.

[2] R. Fahy, J. Van Hoboken, and N. Van Eijk, "Data Privacy, Transparency and the Data-Driven Transformation of Games to Services," *2018 IEEE Games, Entertain. Media Conf. GEM 2018*, pp. 136–146, 2018.

[3] A. R. Pathak, M. Pandey, and S. Rautaray, "Construing the big data based on taxonomy, analytics and approaches," *Iran J. Comput. Sci.*, vol. 1, no. 4, pp. 237–259, 2018.

[4] O. R. D.J. Patil, "Building Data Science Teams." Oreilly Radar, 2011.

[5] A. Immonen and J. Kalaoja, "Requirements of an Energy Data Ecosystem," *IEEE Access*, vol. 7, pp. 111692–111708, 2019.

[6] D. R.-M. and P. T. Michael Schroeck, Rebecca Shockley, Janet Smart, "Analytics : The real-world use of big data." Saïd Business School at the University of Oxford, pp. 1–22, 2012.

[7] Gartner, "What is a big data." Gartner Publications, pp. 1–3, 2019.

[8] J. M. Nobubele, and Angel Shozi, "Big data privacy and security: A systematic analysis of current and future challenges." University of South Africa, South Africa, pp. 1–9.

[9] S. Kumar and M. Singh, "A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem," *Big Data Min. Anal.*, vol. 2, no. 4, pp. 240–247, 2019.

[10] G. F. Government and German Federal Government, "Digital Agenda 2014 - 2017 for Germany," pp. 1–18, 2014.

[11] D. Faggella, "Data Dominance – How Companies and Countries Win with Artificial Intelligence." Emerj Artificial Intelligence Research, pp. 1–12, 2019.

[12] OECD, "Data-driven Innovation for Growth and Well-being." OECD, pp. 1–86, 2014.

[13] S. Allard, "DataONE: Facilitating eScience through Collaboration," *J. eScience Librariansh.*, vol. 1, no. 1, pp. 4–17, 2012.

[14] S. J. Divakar Mysore, Shrikant Khupat, "Introduction to Big Data Architecture," pp. 1–14, 2017.

[15] KnowledgeHut, "Types of Big data." KnowledgeHut, pp. 1–4, 2019.

[16] S. Mazumdar, D. Seybold, K. Kritikos, and Y. Verginadis, *A survey on data storage and placement methodologies for Cloud-Big Data ecosystem*, vol. 6, no. 1. Springer International Publishing, 2019.

[17] L. Ding, V. Peristeras, and M. Hausenblas, "Linked Open Government Data [Guest editors' introduction]," *IEEE Intell. Syst.*, vol. 27, no. 3, pp. 11–15, 2012.

[18] D. Lee, "Building an open data ecosystem," pp. 351–360, 2015.

[19] D. Misra, A. Mishra, S. Babbar, and V. Gupta, "Open Government Data Policy and Indian Ecosystems," pp. 218–227, 2017.

[20] European Commission, "Data Strategy For Digital Transformation." European Commission, pp. 1–22.

[21] Statista organization, "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)." Statista Survey Organization.

[22] B. Xu, L. Da Xu, H. Cai, C. Xie, J. Hu, and F. Bu, "Ubiquitous data accessing method in iot-based information system for emergency medical services," *IEEE Trans. Ind. Informatics*, vol. 10, no. 2, pp. 1578–1586, 2014.

[23] M. M. Rantanen, S. Hyrynsalmi, and S. M. Hyrynsalmi, "Towards Ethical Data Ecosystems: A Literature Study," *Proc. - 2019 IEEE Int. Conf. Eng. Technol. Innov. ICE/ITMC 2019*, 2019.

[24] Kitchenham, "Guidelines for performing Systematic Literature Reviews in Software Engineering." Keele University, UK and University of Durham, UK, pp. 1–57, 2007.

[25] E. Haak, J. Ubacht, M. Van Den Homberg, S. Cunningham, and B. Van Den Walle, "A framework for strengthening data ecosystems to serve humanitarian purposes," *ACM Int. Conf. Proceeding Ser.*, 2018.

[26] N. Raymond, Z. Al Achkar, S. Verhulst, J. Berens, L. Barajas, and M. Easton, "Building data responsibility into humanitarian action," *OCHA Policy Stud. Ser.*, no. May, p. 18, 2016.

[27] M. A. Parsons *et al.*, "A conceptual framework for managing very diverse data for complex, interdisciplinary science," *J. Inf. Sci.*, vol. 37, no. 6, pp. 555–569, 2011.

[28] A. Zuiderwijk, M. Janssen, and C. Davis, "Innovation with open data: Essential elements of open data ecosystems," *Inf. Polity*, vol. 19, no. 1–2, pp. 17–33, 2014.

[29] M. I. S. Oliveira, L. E. R. A. Oliveira, M. G. R. Batista, and B. F. Lóscio, "Towards a meta-model for data ecosystems," pp. 1–10, 2018.

[30] A. Zuiderwijk, M. Janssen, G. Van De Kaa, and K. Poulis, "The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments," *Inf. Polity*, vol. 21, no. 3, pp. 223–236, 2016.

[31] A. Poikola, P. Kola, and K. A. Hintikka, *Public data-an introduction to opening information resources*. 2011.

[32] S. S. Dawes, L. Vidiasova, and O. Parkhimovich, "Planning and designing open government data programs: An ecosystem approach," *Gov. Inf. Q.*, vol. 33, no. 1, pp. 15–27, 2016.

[33] G. Smith, H. A. Ofe, and J. Sandberg, "Digital service innovation from open data: exploring the value proposition of an open data marketplace," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2016-March, pp. 1277–1286, 2016.

[34] M. Van Den Homberg, J. Visser, and M. Van Der Veen, "Unpacking data preparedness from a humanitarian decision making perspective: Toward an assessment framework at subnational level," *Proc. Int. ISCRAM Conf.*, vol. 2017-May, no. May, pp. 2–13, 2017.

[35] C. Moiso and R. Minerva, "Towards a user-centric personal data ecosystem the role of the bank of individuals' data," *2012 16th Int. Conf. Intell. Next Gener. Networks, ICIN 2012*, pp. 202–209, 2012.

[36] M. I. S. Oliveira, G. de F. B. Lima, and B. F. Lóscio, "Investigations into Data Ecosystems: a systematic mapping study," *Knowl. Inf. Syst.*, pp. 1–42, 2019.

[37] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," in *2014 International Conference on Collaboration Technologies and Systems (CTS)*, 2014, pp. 104–112.

[38] Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data," *SIGMOD-PODS'09 - Proc. Int. Conf. Manag. Data 28th Symp. Princ. Database Syst.*, pp. 1005–1009, 2009.

[39] G. Pal, G. Li, and K. Atkinson, "Big Data Real Time Ingestion and Machine Learning," *Proc. 2018 IEEE 2nd Int. Conf. Data Stream Min. Process. DSMP 2018*, pp. 25–31, 2018.

[40] K. Lam, A. Kwan, and K. Ramamritham, "RTMonitor : Real-Time Data Monitoring Using Mobile Agent Technologies," *Proc. 28th Int. Conf.*

Very Large Data Bases, 2002.
[41] R. Wang, T. F. Edgar, M. Baldea, M. Nixon, W. Wojsznis, and R. Dunia, "A geometric method for batch data visualization, process monitoring and fault detection," *J. Process Control*, vol. 67, pp. 197–205, 2018.
[42] P. Muthulakshmi and S. Udhayapriya, "a Survey on Big Data Issues and Challenges," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 6, pp. 1238–1244, 2018.
[43] U. Aftab and G. F. Siddiqui, "Big Data Augmentation with Data Warehouse: A Survey," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 2785–2794, 2019.
[44] United Nations, "New data sources for official statistics – access, use and new skills." United Nations - Economic and Social Council, pp. 1–19, 2019.
[45] S. Fan, R. Y. K. Lau, and J. L. Zhao, "Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix," *Big Data Res.*, vol. 2, no. 1, pp. 28–32, 2015.
[46] A. U. Khan and B. K. Ratha, "Business data extraction from social networking," *2016 3rd Int. Conf. Recent Adv. Inf. Technol. RAIT 2016*, pp. 651–656, 2016.
[47] E. Przeybilovicz and M. A. Cunha, "Open government data programs," in *Proceedings of the 19th Annual International Conference on Digital Government Research Governance in the Data Age - dgo '18*, 2018, pp. 1–2.
[48] M. Kuneva, "Key note speech in an event 'Roundtable on Online Data Collection, Targeting and Profiling.'" EU official Website, pp. 1–4, 2009.
[49] E. Biliri *et al.*, "Big data analytics in public safety and personal security: Challenges and potential," *2017 Int. Conf. Eng. Technol. Innov. Eng. Technol. Innov. Manag. Beyond 2020 New Challenges, New Approaches, ICE/ITMC 2017 - Proc.*, vol. 2018-Janua, pp. 1382–1386, 2018.
[50] European Commission, "Data Ecosystem Architectural Model Document." European Commission, pp. 1–164.
[51] A. Immonen, M. Palviainen, and E. Ovaska, "Requirements of an open data based business ecosystem," *IEEE Access*, vol. 2, pp. 88–103, 2014.
[52] European Commission, "EU Open Data Portal." EU official Website, pp. 1–4, 2019.
[53] European Commission, "European Commission Digital Strategy." European Commission, pp. 1–33, 2018.
[54] H. A. Smith and J. D. McKeen, "Developments in Practice XXX: Master Data Management: Salvation Or Snake Oil?," *Commun. Assoc. Inf. Syst.*, vol. 23, 2008.
[55] F. Haneem, R. Ali, N. Kama, and S. Basri, "Resolving data duplication, inaccuracy and inconsistency issues using Master Data Management," *Int. Conf. Res. Innov. Inf. Syst. ICRIIS*, pp. 1–6, 2017.
[56] E. Baghi, S. Schlosser, V. Ebner, B. Otto, and H. Oesterle, "Toward a decision model for master data application architecture," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 3827–3836, 2014.
[57] Gartner, "IT Glossary - Master Data Management - MDM." Gartner Publications, pp. 1–7, 2019.
[58] B. Otto, K. M. Hüner, and H. Österle, "Toward a functional reference model for master data quality management," *Inf. Syst. E-bus. Manag.*, vol. 10, no. 3, pp. 395–425, 2012.
[59] Michigan University-USA, "What is metadata." Michigan University - USA, pp. 1–4, 2019.
[60] Jenn Riley, "Understanding M Etadata What Is Metadata ,." National Information Standards Organization (NISO), 2017.
[61] P. W. Alexander Borek, Ajith K. Parlikad, Jela Webb, *Total Information Risk Management- maximizing the value of data and information assets*. Front Matter, 2014.
[62] D. S. R. Group, "Overview of the DDI Version 3 . 0 Conceptual Model." Structural Reform Group, 2004.
[63] W. K. Michener and M. B. Jones, "Ecoinformatics: Supporting ecology as a data-intensive science," *Trends Ecol. Evol.*, vol. 27, no. 2, pp. 85–93, 2012.
[64] M. El Arass, I. Tikito, and N. Souissi, "An audit framework for data lifecycles in a big data context," *2018 Int. Conf. Sel. Top. Mob. Wirel. Networking, MoWNeT 2018*, no. June, 2018.
[65] M. El Arass, I. Tikito, and N. Souissi, "Data lifecycles analysis: Towards intelligent cycle," *2017 Intell. Syst. Comput. Vision, ISCV 2017*, 2017.
[66] O. Catteau, P. Vidal, and J. Broisin, "A generic representation allowing for expression of learning object and metadata lifecycle," *Proc. - Sixth Int. Conf. Adv. Learn. Technol. ICALT 2006*, vol. 2006, pp. 30–32, 2006.
[67] R. L. Grossman, "Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data," *Trends Genet.*, vol. 35, no. 3, pp. 223–234, 2019.
[68] L. Ding *et al.*, "TWC LOGD: A portal for linked open government data ecosystems," *J. Web Semant.*, vol. 9, no. 3, pp. 325–333, 2011.
[69] F. A. Zeleti and A. Ojo, "Open data value capability architecture," *Inf. Syst. Front.*, vol. 19, no. 2, pp. 337–360, 2017.
[70] Y. Chen, S. Chen, and S. C. Lin, "A metadata lifecycle model for digital libraries: methodology and application for an evidence-based approach to library research," *Computing*, no. August, pp. 1–15, 2003.
[71] S. M. Lee and S. Hong, "An enterprise-wide knowledge management system infrastructure," *Ind. Manag. Data Syst.*, vol. 102, no. 1, pp. 17–25, 2002.
[72] Alex Ball, "Review of Data Management Lifecycle Models." University of Bath, p. 15, 2012.
[73] K. Möller, "Lifecycle models of data-centric systems and domains The abstract data lifecycle model," *Semant. Web*, vol. 4, pp. 67–88, 2013.
[74] S. McKeever, "Understanding web content management systems: Evolution, lifecycle and market," *Ind. Manag. Data Syst.*, vol. 103, no. 8–9, pp. 686–692, 2003.
[75] IBM, "Wrangling big data: Fundamentals of data lifecycle management," *IBM Manag. data lifecycle*, 2013.
[76] Research Data Management Team, "Data Life Cycle & Data management planning." University of Essex, UK, 2013.
[77] J. L. Faundeen *et al.*, "The United States Geological Survey Science Data Lifecycle Model," *U.S. Geological Survey Open-File Report 2013–1265*. USA Govt., pp. 1–4, 2013.
[78] N. Khan *et al.*, "Big data: Survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, pp. 1–19, 2014.
[79] M. El Arass and N. Souissi, "Data Lifecycle: From Big Data to SmartData," *Colloq. Inf. Sci. Technol. Cist*, vol. 2018-Octob, no. October, pp. 80–87, 2018.
[80] A. Michota and S. Katsikas, "Designing a seamless privacy policy for social networks," *ACM Int. Conf. Proceeding Ser.*, vol. 01-03-Octo, pp. 139–143, 2015.
[81] Ministry of Education, "Action Plan - Implementing DataStrategy@EC." European Commission, pp. 1–18, 2018.
[82] D. C. C. Curation and L. Model, "DCC Curation Lifecycle Model Key elements of the DCC Curation Lifecycle Model," pp. 2–4, 2015.
[83] A. Burton and A. Treloar, "The International Journal of Digital Curation Designing for Discovery and Re-Use : the ' ANDS Data Sharing Verbs ' Approach to Service Decomposition," *Discovery*, vol. 4, no. 3, pp. 44–56, 2009.
[84] K. Crowston and J. Qin, "A capability maturity model for scientific data management: Evidence from the literature," *Proc. ASIST Annu. Meet.*, vol. 48, no. 1, pp. 1–22, 2011.
[85] NIST Big Data Public Working Group: Definitions and Taxonomies Subgroup, "NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies," vol. 2. NIST, p. 31, 2015.
[86] E. Lee, "Categories and Subject Descriptors," *Proc. 47th Des. Autom. Conf. (DAC),ACM*, pp. 737–742, 2010.
[87] H. D. A. do. Santos, M. I. S. Oliveira, G. de F. A. B. Lima, K. M. da Silva, R. I. V. C. Rayelle, and B. F. Lóscio, "Investigations into data published and consumed on the Web: a systematic mapping study," *J. Brazilian Comput. Soc.*, vol. 24, no. 1, 2018.
[88] P. O. T.D., "Public service Data Strategy 2019-2023." Government of Ireland, pp. 1–40, 2018.
[89] L. P. Jennifer L. Bauer and and Haley St. . Dennis, "Data Brokers and Human Rights (Big Data, Big Business)," *SSRN Electronic Journal*. Institute for Human Rights and Business (IHRB), pp. 1–24, 2016.
[90] S. Orenga-Roglá and R. Chalmeta, "Framework for implementing a big data ecosystem in organizations," *Commun. ACM*, vol. 62, no. 1, pp. 58–65, 2019.
[91] L. Lin, T. Liu, J. Hu, and J. Zhang, "A privacy-aware cloud service selection method toward data life-cycle," *Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS*, vol. 2015-April, pp. 752–759, 2014.
[92] S. Bouteiller, "How to manage corporate data to create value - CIGREF." CIGREF, pp. 1–64, 2014.
[93] M. Kaufmann, "Towards a Reference Model for Big Data Management.," 2016.
[94] L. M. Ambrosio *et al.*, "An approach to support data integration in a scientific software ecosystem platform," *Proc. 2019 IEEE 23rd Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2019*, pp. 39–44, 2019.
[95] P. Cuffe, "Data Visualization: A Practical Introduction," *IEEE Trans. Prof. Commun.*, vol. 62, no. 3, pp. 310–311, 2019.
[96] B. Lóscio BF, Oliveira MIS, "Web Publishing and Consumption: Concepts and Challenges." SBBD, pp. 1–95, 2015.
[97] M. J. Becker, "The consumer data revolution: The reshaping of industry competition and a new perspective on privacy," *J. Direct, Data Digit. Mark. Pract.*, vol. 15, no. 3, pp. 213–218, 2014.

[98] S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. Ben Yahia, and D. Draheim, "Towards Disaster Resilient Smart Cities: Can Internet of Things and Big Data Analytics Be the Game Changers?," *IEEE Access*, vol. 7, pp. 91885–91903, 2019.
[99] Information Builders, "Achieving Digital Transformation Using IoT and Big Data." Information Builders, pp. 1–14, 2017.
[100] OECD, "Going digital in a multilateral world," *Oecd*, no. May. OECD, pp. 1–106, 2018.
[101] R. L. Grossman, "data lakes, commons and clouds : A review for analyzing and sharing genomic data." Arxiv, pp. 1–5, 2018.
[102] J. J. Zubcoff *et al.*, "The university as an open data ecosystem," *Int. J. Des. Nat. Ecodynamics*, vol. 11, no. 3, pp. 250–257, 2016.
[103] J. Attard, F. Orlandi, and S. Auer, "Data Value Networks: Enabling a New Data Ecosystem," *Proc. - 2016 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2016*, pp. 453–456, 2017.
[104] T. M. Harrison, T. A. Pardo, and M. Cook, "Creating Open Government Ecosystems: A Research and Development Agenda," *Futur. Internet*, vol. 4, no. 4, pp. 900–928, 2012.
[105] Adner R. and Kapoor R., "Value Creation in Innovation Ecosystems: How the Structure of Technological Interdependence Affects Firm Performance in New Technology Generations," *Strateg. Manag. J.*, vol. 31, no. 1, pp. 306–333, 2010.
[106] H. B. James E. Moore, "The Death of Competition -Leadership and Strategy in the Age of Business Ecosystems." New York: Harper Business, pp. 1–6, 1996.
[107] M. I. and R. Levien, *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability*. Harvard Business School Press, 2004.
[108] R. Adner, *Match your innovation strategy to your innovation ecosystem,*. Harvard Business School Press, 2006.
[109] M. L. Brodie, "Data Integration at Scale: From Relational Data Integration to Information Ecosystems," *2010 24th IEEE Int. Conf. Adv. Inf. Netw. Appl.*, pp. 1–3, 2010.
[110] A. B. H. S. Mhamdia, "Performance measurement practices in software ecosystem," *Int. J. Product. Perform. Manag.*, vol. 62, no. 5, pp. 514–533, 2013.
[111] M. Hadzic and T. S. Dillon, "Application of digital ecosystems in health domain," *2008 2nd IEEE Int. Conf. Digit. Ecosyst. Technol. IEEE-DEST 2008*, pp. 543–547, 2008.
[112] M. Heimstädt, "'The Institutionalization of Digital Openness,'" in *Proceedings of The International Symposium on Open Collaboration - OpenSym '14*, 2014, pp. 1–2.
[113] V. Köster and G. Suárez, "Open data for development: Experience of Uruguay," *ACM Int. Conf. Proceeding Ser.*, vol. 01-03-Marc, pp. 207–210, 2016.