Journal of Big Data

RESEARCH Open Access

DaLiF: a data lifecycle framework for data-driven governments

Syed Iftikhar Hussain Shah[1*], Vassilios Peristeras[1,2] and Ioannis Magnisalis[1,3]

*Correspondence:
i.shah@ihu.edu.gr
[1] School of Science and Technology, International Hellenic University, Thessaloniki, Greece
Full list of author information is available at the end of the article

Abstract

The public sector, private firms, business community, and civil society are generating data that is high in volume, veracity, velocity and comes from a diversity of sources. This kind of data is known as big data. Public Administrations (PAs) pursue big data as "new oil" and implement data-centric policies to transform data into knowledge, to promote good governance, transparency, innovative digital services, and citizens' engagement in public policy. From the above, the Government Big Data Ecosystem (GBDE) emerges. Managing big data throughout its lifecycle becomes a challenging task for governmental organizations. Despite the vast interest in this ecosystem, appropriate big data management is still a challenge. This study intends to fill the above-mentioned gap by proposing a data lifecycle framework for data-driven governments. Through a Systematic Literature Review, we identified and analysed 76 data lifecycles models to propose a data lifecycle framework for data-driven governments (DaliF). In this way, we contribute to the ongoing discussion around big data management, which attracts researchers' and practitioners' interest.

Keywords: Big data, Data-driven government, Government Big Data Ecosystem, Data management, Data lifecycle

Introduction

Big data is sweeping across numerous fields of public and private organizations. Becker [1] highlights that "big data is the oil of the 21st century" as the capability to exploit big data has become a vital success factor for PAs, private firms, and civil society.

Recently, data-driven public and private organizations define data policies and draft data strategies connected with their organization's vision, mission, and goals. They also attempt to secure expertise and skills in data-related areas, put in place data technologies in areas covering data generation, enrichment, storage, access, sharing, publishing, management, analysis, use, protection, privacy, and archive [2–4].

Data-driven transformation for large organizations is a rigorous, resource-consuming, and long-term undertaking that influences technology, people, organizations, and cultures [5]. Data champions widely use big data and big data analytics in every aspect of their businesses, which include sales, marketing, supply chain, manufacturing, R&D, and HR management [2–4].

In the public sector, the implementation of big data tools, technologies and tactics offers opportunities including effective public service delivery, evidence-based decision making for policymakers, growth in the digital economy, creation of new professional jobs, encouragement of civic participation in the definition and enhancement of public policies [6–9].

A Big Data Ecosystem (BDE) is a complex set of various interconnected components related to big data, models, organisational structures, and roles covering the whole data lifecycle [10]. It consists of diverse components, including data infrastructure, data models, data analytics, as well as organisational and cultural components [10–12].

Leveraging data to provide the most public value rests on PA's capability to cultivate and sustain a useful and effective data management environment. While achieving this requires various functions, roles, responsibilities, abilities, and skills for both technology and people, it is critical to pay special care to the data life cycle [13]. The data lifecycle represents all phases during its life, from its planning, collection to its distribution, use, and reuse, and destruction [14]. The data life cycle provides a high-level overview of the phases involved in the successful management of big data for its use and reuses [6, 15]. Data lifecycle is helpful to identify dataflows and work processes for stakeholders in the GBDE. Moreover, identifying the data lifecycle that fits a company's data usage is a critical task for the organizations, and the way the data lifecycle is managed is also important to transform data into knowledge [16] and to extract value from big data to improve organization data operations [17].

In this study, we proposed a DaLiF. We conducted a Systematic Literature Review (SLR) [18–20].

The remainder of the paper is structured as follows. In Section 2, we mention the background of our research work. Section 3 explains our research methodology. Section 4 presents the results of the literature review and research implications. In the last section, we illustrate the conclusion, the limitations of this study, and propose future work.

Background and scope of this work

This section illustrates the background of this research, BDE fundamental theories, identified research gaps that have attracted our attention, and our contribution.

Overview of the Government Big Data Ecosystem and data lifecycle fields

The word "data" originates from the Latin phrase 'datum' [5]. Data is a discrete, limitless entity that has an unstructured and unprocessed shape. Organizations further process such kinds of data, as per their needs, to illustrate relevant objects, events, concepts, or facts [5].

Big data is a concept that is characterized by data that have high volume, veracity, velocity, and variety. Big data needs economical, advanced ways of information processing to be used for generating insight and supporting decision making [21–23]. In data-driven organizations, big data is regarded as a critical strategic asset [24]. The capability to manage big data generates opportunities for organizations to attain viable benefits in the present digitized marketplace [25, 26]. This capability requires cost-effective and distinctive novel big data tools to be put in place [22, 26, 27]. The exploitation of big data signifies a paradigm shift in tactics to comprehend and study the world [6]. Private and

public organizations are currently flooded with a massive quantity of big data produced with high speed [5] through diverse and "smart" data sources. Such big data sources include people, the Internet, smart mobile handset, online social networks (Twitter, Facebook, LinkedIn, Instagram), the Internet of Things (IoT), autonomous vehicle, Global Positioning Systems, smart cities, etc. [28–35]. Globally, there are over 3.5 billion social network users and about 26.66 billion IoT linked devices and sensors [8, 32]. It is forecasted that about 75 billion IoT devices will be connected, and between 100 million and 2 billion human genomes will be sequenced by 2025 [36]. The generated data relate to all public sectors, e.g., health, agriculture, manufacturing, justice, transportation, education, and social welfare [8, 31, 32].

We adopt the "ecosystem lens" for exploring this phenomenon. This view is useful to understand the interdependencies among collaborators in exchange networks [37, 38]. The BDE reveals a complex, connected ecosystem of high-capacity networks, data users, data applications, and services required to filter, store, archive, process, share, and visualize data that is gathered from multiple data sources [39, 40].

The BDE stakeholders consist of public and private organizations, civil society, development partners, and users. This ecosystem can assist PAs to promote evidence-based decision making, secure data interoperability, data security, and privacy protection, prioritize requirements and challenges, to promote civic participation, and finally contributes to effective governance [7–9]. All the above signify the Government Big Data Ecosystem (GBDE).

Managing big data during its lifecycle model is a collection of challenging tasks with the objective to extract value [6, 17, 41].

In GBDE, a data lifecycle model is a key data management tool for organizations [14, 15] Data management intends to provide data, which is complete, precise, readable, and accessible to data users. A data lifecycle provides a high-level framework to plan, organize and manage all aspects of data during its life phases, from data planning, collection to its destruction, and the relationship between phases [6, 14, 14, 15, 41–45]. Blazquez, and Domenech highlighted that a data lifecycle is the series of stages that data follow from the moment they enter a system to the moment they are deleted from the system or stored [46]. Shameli-Sendi stated that a data life cycle entails where the data is generated, where it is modified, processed, where it is transmitted, where it is presented, and where it will finally be put in storage [47]. Moreover, data go through various phases that may vary depending on the kind of data and goal to be achieved. Each phase of the data lifecycle offers certain functions and contributes to better big data management. Big data undergoes various phases like data collection, integration, analysis, publication, and destruction by different actors for numerous purposes. Such set of phases in combination constitute a data lifecycle [48–50]. Examples of data lifecycles in the literature include the IBM data lifecycle [51], the open government data lifecycle [52], the DataONE lifecycle [15], the Abstract Personal Data Lifecycle [48], the Research data lifecycle [53], etc. We present and analyse the literature data lifecycle models in the forthcoming sections of the paper.

There are numerous benefits deriving from designing and implementing in a consistent way a data lifecycle for PAs. These benefits include, but not limited to the followings: (i) ease in planning and handling complexity of data management in all data life phases

[15, 42–44, 53–55], (ii) identifying and illustrating a sequence of all essential activities related to data, (iii) support organizations for the preparation of data products for the data users [42–44, 54, 55], (iv) help data users to have a well understanding of the data assets available to them [56], (v) effective gathering of data including metadata from various (internal and external) sources [53, 57, 58], (vi) implementation of the once-only principle [59], (vii) creation of a homogeneous set of data through consolidation [6, 60], (viii) identify, remove noise, uncertainty, and errors in collected data, and maintain data quality [56, 61, 62], (ix) addition of appropriate data for completion and improvement [61, 63], (x) better analysis of data to extract knowledge and discover new insights so that policymakers use this knowledge to generate desire value [42–44, 61] (xi) visualize data for a better understanding of a common person and its usage for future course of actions [58, 64], (xii) support to adopt appropriate data storage approach to ensure the data availability and scalability [15, 65], (xiii) assistance to promote the use of data with the consent of the owner of data [66, 67], (xiv) create an opportunity to the stakeholders to offer their viewpoints on the data [49, 52], (xv) aid PAs to ensure the protection of big data, including personal data, and promote effective governance [62, 68, 69], (xvi) support organizations to manage big data throughout the lifecycle [6, 41], and (xvii) help software designers to create sustainable software for big data management [42–44, 54].

Industry Standards for Data management We studied Data management—Body of Knowledge (DM-BOK) offered by DAMA- Data Management Association. We used DM-BOK, as an industry standard for data management, in this study. DMBOK is a comprehensive data-oriented framework compared to TOGAF - The Open Group Architecture Framework, and COBIT- Control Objectives for Information and Related Technologies [70]. The TOGAF defines the process for creating a data architecture as part of overall enterprise architecture, and it can be a precursor to implementing data management. In comparison, COBIT provides data governance as part of overall IT governance. It can offer a Maturity Model for assessing data management. There are significant differences in their scopes; however, there are some commonalities in terms of data management, like the data governance concept.

DMBOK manages data across the entire lifecycle. It offers a detailed framework to support the development and implementation of data management processes and procedures [70–72]. DMBOK consists of ten functions and activities inside. The ten functions include Data Governance, Data Architecture Management, Data Development, Data Operations Management, Reference and Master Data Management, Data Security Management, Data Warehouse and Business Intelligence Management, Document and Content Management, Meta-Data Management, and Data Quality Management. Each DMBOK functions comprise data management environmental elements like goals and principles, roles and responsibilities, practices, and techniques. However, such data management frameworks are very generic in the guidelines [70, 73].

BDE fundamental theories

A theory is a presumption or a system of ideas aimed to explain reality. It offers a well-substantiated explanation about an aspect of the real world [74]. A valuable theory provides us a framework for raising various questions about something [11, 74, 75]. The existing research studies present a heterogeneous theoretical foundation to define BDEs

and related aspects. Such theories are often influenced by socio-technical, ecosystems, platform, actor-network, business process management, and value chain theories. These mixed theories are usually used in the literature to cover the theoretical gap as the big data field is in the early stages [11, 76]. Numerous business, research, and industry communities study the big data field [76–78]. For example, in the case of definitions of BDE, some definitions stay relevant to specific domains like humanitarian [79–81], and personal data ecosystems [67]. Such studies have a narrow perspective, focus on a specific notion with partial details, and describe BDE definitions and other related terminologies that vary considerably [81–83]. We observe similar issues in the case of data lifecycles studies as well. This is due to the fact that the BDE area is in its infancy, and different research and business communities have been investigating the area separately [76]. Therefore, currently, the existing BDE theory does not offer a full conceptual foundation for further studies into the research field. To extend insights into the current state of the BDE, data lifecycle, and other related aspects, we conduct this holistic SLR as a theoretical groundwork about BDE.

Research gaps

We identify the following research gaps while studying the data lifecycle for GBDE.

In the literature, we found several specific areas/domains where data lifecycles have been proposed including 'scientific research' [55, 84, 85], 'media production' [57], Semantic Web [49], databases [86], open data [28], information systems [87], and cloud computing [88]. Furthermore, we found certain data lifecycle studies focusing on specific big data aspects. Such data aspects include data quality [17, 55], data processing [28], data conception [89, 90], data management [55], security [88], data strategy [30, 66], analytic [91], feedback and data refinement [92, 93]. In the literature, the various data lifecycles exhibit many differences particularly about the objectives, intended audience, phases, actors, and attention [94].

We also observed that the proposed phases of the data lifecycles also vary [56] while often for the same phase, different terms are used. For example, in the case of data collection, some studies use the term "generation" [95, 96] while in other data lifecycles call it "receive" [17, 90], "acquir" [55, 63, 73, 97], or data "capture" [84]. So, there is a lot of confusion for the research community and practitioners to understand and use such phases and associate them with management practices in their organizations. So far, we could not find an all-inclusive data lifecycle model for GBDE.

In the literature, we did not find many studies that attempt to align their data lifecycles with industry-standard for data management like DM.BOK.

Our contribution

In this research article, we mainly concentrate on addressing the above-mentioned literature research gaps by suggesting an all-inclusive big data lifecycle for data-driven governments. We call this DaLiF. We found and analysed 76 data lifecycle models published during the last 25 years. We provide our research approach, a detailed description of the literature for data lifecycles, and DaLiF in the forthcoming sections of the study. The explanation about what we add on the top of our proposed data lifecycle for GBDE is as below:

(i) It can be applied in various public sector areas like health, agriculture, education while considering these areas related to public organizations' business processes, requirements, and environment. (ii) This data lifecycle type is evolving; i.e., there is no requirement for government big data to go through the whole lifecycle before a fresh iteration can be commenced. Moreover, DaLiF phases can be passed in several different orders and, in principle, for an infinite number of times. (iii) DaLiF contains phases that can contribute to creating Open Government (OGD's) values, namely politic, economic, and social values as proposed in [98] as future research work. (iv) It is the first big data lifecycle consists of fourteen phases, which were reported as missing phases in [17, 61, 98], and this data lifecycle corresponds to the public administration vision about data management in GBDE. (v) Specific functions at each phase prove proposed lifecycle completeness with respect to the big data 4Vs challenges, i.e., Volume, Variety, Velocity, and Veracity. (vi) We consider the data protection phase that exists at each phase of the data lifecycle to tackle major concerns of big data like privacy, integrity, availability, confidentiality, and data security protection issues in GBDE. (vii) The mining of valuable data from a large influx of information is a critical issue in big data.Therefore, we include data enrichment phase in DaLiF to enrich data to qualify and validate the related aspects in GBDE. (viii) DaLiF contains data quality' phase along with key functions, like conformance to data quality business rules, in order to ensure high data quality. (ix) We consider data quality, protection, storage, archive at each phase of the data lifecycle for appropriate data management in GBDE. (x) We categorically focus on the data governance phase, which is the overall process of managing and controlling government big data to maximize data usage and contribute to value creation in GBDE. (xi) Finally, we mapped the DaLiF with DM-BOK, which is an industry-standard for data management.

Research method

In the earlier section, we found the research gaps while reviewing the GBDE, including data lifecycles. This study aims to mitigate the above-mentioned research gaps by identifying relevant exiting data lifecycle models, performing analysis of these literature models, and proposing the DaLiF based on the existing data lifecycles models.

To accomplish the above-mentioned research goal, we performed qualitative research about the government big data ecosystem and data lifecycles by conducting a Systematic Literature Review (SLR). SLR is a research approach to find, evaluate, and explain research work, literature created by scholars, researchers, and practitioners [99]. Fink described the literature review as a systematic, specific, and reproducible method to find, evaluate, and synthesize the existing research work delivered by researchers, scholars, and practitioners [99]. We practiced this research methodology following guidelines from the literature [18–20]. We describe our approach in five steps. The SLR process or research review protocol's *first step* is focused on devising the research questions. The *second step* mainly concentrates on three sub-activities: selecting digital research libraries, creating search strings, and literature search. The *third step* is primarily focused on identifying the relevant research articles and applying quality assessment to the articles. It is supported by specifying and applying inclusion and exclusion criteria. The *fourth step* intends to examine the research articles, mine relevant information, perform

verification of results, and connect the findings to research gaps. The *last step* describes the research outcomes and organizes them in our proposed data lifecycle for the GBDE.

The Fig. 1 summarises the steps. We briefly present them in the subsequent sections.

Goal and research questions

Our research aims to identify and analyse the existing data lifecycle models, find common and complementary phases with their distinct labels of data lifecycles, and suggest an extensive data lifecycle framework for data-driven governments. For this, we outline the Research Questions.

To produce the research questions, we conducted a preliminary review of the related literature about GBDE. We identified the above-mentioned research gaps related to the data lifecycles and used these gaps to focus on our research questions. We implemented a combination of the following two Basic gap-spotting modes: confusion spotting and application spotting, as an approach to construct our research questions [100, 101]. We partially implemented the PEO framework [101–103] to give appropriate structure to our research questions to ensure clarity. Additionally, we evaluated the robustness of our qualitative research question by using FINER criteria [103–105] to ensure that our research questions are feasible, interesting, novel, ethical, and relevant.

As a result of the research review protocol's step 1, "formulating the research questions", we devised the following research questions.

(RQ1) What are the existing data lifecycle models described and the phases they introduce in the literature?

(RQ2) What is an all-inclusive big Data Lifecycle Framework for data-driven governments and its phases?

RQ1 aims to find and organise the literature for the data ecosystem lifecycle. *RQ2* intends to propose a new big data lifecycle framework for data-driven governments and

Fig. 1 Steps in the SLR process

its phases, we call this DaLiF, which considers and builds on top of the current state-of-the-art data life cycle models.

We incorporate a comprehensive analysis of literature data lifecycle models and explain DaLiF and its phases. We showcase a graphical view of DaLiF and deliberate our research findings.

Finding relevant research work

SLR uses a comprehensive method to review the findings presented in prior published research. Literature research aims to carry out a systematic review that demands extensive coverage of current research on the research topic of interest within the stipulated period. Most researchers endeavored to work out a systematic review opinioned in a research survey that they did not stop off their literature search action until they believed they had attained their target [19]. This section explains the above-mentioned SLR process second step that centers on offering the following details about our selected digital research libraries, the procedure for the discovery of search strings, and the searching process.

Selection of digital research libraries To carry on with the SLR, the selection of digital research libraries is the phase when researchers make a decision about where to search and how to search for required studies that contain relevant information for the research questions of the study [106].

As a result of this sub-step, We selected and examined the following four digital research libraries for the literature search, ACM, Science Direct, IEEE Xplore, and Springer Link.

Formulation of search string We devised search strings to discover research articles from the above-mentioned selected digital research libraries based on the following actions:

1. We formulated search strings related to the above-mentioned RQs.
2. In the case of the critical aspect, like data actor, we find alternate words and synonyms for these keywords.
3. We use Boolean operators like 'OR', 'AND' to extend the search by adding other words and synonyms.
4. In some cases, we also adjust/alter the search string.
5. Each search string includes one or more than one keywords like "data lifecycle" or "data model".

The above-mentioned measures are applied to formulate search strings for our research questions. As a result of this sub-step, we formulated and used different search strings which are mentioned in Additional file 1 to meet the journal's page limits.

Literature search We began the literature search to obtain relevant research papers in February 2019 and carried out this procedure until May 2021. All results from searches are primarily based on titles, keywords, and abstracts. We applied the above-mentioned measures to formulate search strings for our literature search process. To get relevant literature about DaLiF, we have completed a literature search activity in the following two stages.

Stage-I: We used the aforementioned four digital libraries to search strings and their variants, which are based on the following keywords:- "DATA LIFECYCLE", "DATA LIFE CYCLE", "DATA ECOSYSTEM", "BIG DATA LIFECYCLE", "BIG DATA LIFE CYCLE", "GOVERNMENT DATA ECOSYSTEM", "DATA LIFECYCLE FOR GOVERNMENT", "DATA-DRIVEN GOVERNMENT, "DATA LIFECYCLE FOR DATA-DRIVEN GOVERNMENT", "DATA LIFECYCLE FOR PUBLIC ADMINISTRATION" along with choices "exact phrase" and "matches all". We examined the outcomes of the above-mentioned first stage and matched the results with our crucial research sub-topics regarding big data lifecycle of GBDE. We observed that we require additional relevant research papers, and then we decided to perform the following stage-II.

Stage-II: In this stage, we expanded the search queries performed in stage-I by adding "matches any" instead of options "exact phrase" and "matches all".

Result of SLR process step 2 As a result of the research review protocol's step 2, "finding relevant research work", in total, we collected 1217 research articles. We kept the literature search outcomes in a spreadsheet where every row correlate to a research paper. We recorded various attributes and metadata per paper like paper ID, authors, title, source, keywords, authors, abstract, year of publication, unique viewer tag, searching date, associated search term, and study goal.

Identifying preliminary studies and article quality assessment

The procedure to find preliminary studies, utilizing inclusion and exclusion criteria, and implement a quality assessment is based on the following measures. We rigorously pursued our inclusion and exclusion criteria, given below, to evaluate the relevance of the studies with our research objectives. We manually completed all the next steps to identify preliminary studies. Consequently, we achieved a detailed scrutiny process in the following three phases. We describe our inclusion and exclusion criteria, and then we explain the three phases of our process.

Inclusion criteria To select only the relevant research articles, we included resources that fulfill one or more of the following criteria:

1. Discuss big data lifecycle.
2. Focus on big data lifecycle.
3. Discuss phases of the data lifecycle.
4. Focus on phases of the data lifecycle.
5. Discuss data lifecycle for governments.
6. Focus on data lifecycle for governments.
7. Focus on data lifecycle for GBDE.
8. Resources publication year range is not restricted and keeps it open.
9. Depict applicable results outside of the study.

Exclusion criteria We utilized the following exclusion criteria to filter out research articles which:

1. Are not written in English.
2. Have no relevance to the central theme of the research questions.

3. Have no primary focus on the data lifecycle for GBDE as per the aim of this study.
4. Do not cover the topic of data lifecycle for GBDE.

Scrutiny process In this section, we describe the following key phases to scrutinize the research articles along with the quality assessment of the studies.

Remove duplication We merged and retained the research articles that were found in the preceding literature search phase in a single common folder and removed the research papers that are duplicates. From the above-mentioned literature search process, we gathered a total of 1217 research articles. In this stage, we removed duplications from our study dataset. It decreased the papers to a total of 1198 research articles.

Initial scrutiny based on abstract and title Initially, we investigated the research articles based on abstract and title. If a research paper was not judged for inclusion or exclusion based on these traits, it was added for the next step of examination. Next, titles and abstracts were independently assessed by two researchers. Each researcher noted the research articles that have some uncertainty to decide about the research article's inclusion or exclusion for the next scrutiny phase. Both researchers mostly found similar outcomes, and there was negligible disagreement about the inclusion and exclusion of papers between them. However, both researchers held meetings to settle differences and to discuss disputed and marginal research articles. As a result of the above-mentioned initial scrutiny, we reduced the papers to a total of 578 out of 1198 research articles.

Scrutiny based on the full text In this phase, the research team examined the full text of articles that are already agreed upon in the above-mentioned initial scrutiny phase. The same researchers comprehensively studied and analyzed the full text of the studies. While the third researcher validated and verified the outcomes. The researchers assessed the quality of the research articles based on inclusion and exclusion criteria. We discovered satisfactory quality assessment outcomes in the scrutiny process based on the above-mentioned method that includes, but is not limited to, strict implementation of inclusion and exclusion criteria, internal meetings to resolve minor variances between the researchers, and validation of the results. Moreover, we concentrated on the different vital factors like selection and assessment bias, related to threats to validity as well. The execution of this phase decreased the papers to a total of 232 out of 578 studies. Thus, our preliminary studies include 232 articles.

Result of SLR process step 3 As a result of the research review protocol's step 3, "identifying preliminary studies and applying inclusion and exclusion criteria", our preliminary studies include 232 articles.

We present our literature search strategy results in Fig. 2. We add up research articles of our preliminary studies in a reference manager tool to access, use, and manage references in our research work.

Analysis

In this section, we describe the fourth step of our methodology, i.e., the required data extraction from the research articles and presentation and organization of findings that propose DaLiF to "fill-in" the above-mentioned research gaps.

We comprehensively examined the research articles from our preliminary studies. We mined relevant information that includes existing data lifecycles, phases of data

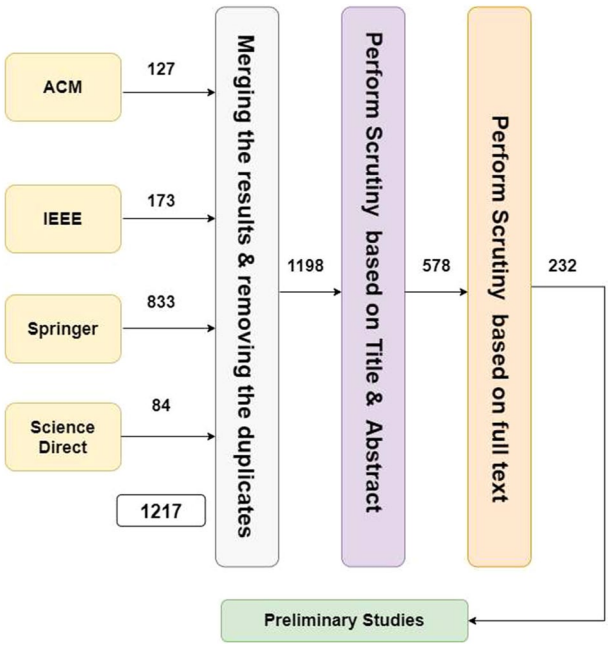

Fig. 2 Procedure for the identification of preliminary studies

lifecycles, objectives, domains, and nature of these data lifecycles. Subsequently, we gathered and classified the mined information and relevant research articles to answer the research questions RQ1-2. We utilized a spreadsheet program as a data extraction template to capture and record the studies' information.

We applied the following steps to extract the outcomes. First, we acquired the general information, like authors, publication year, title, and publication type. Second, studies were examined according to the above-mentioned inclusion and exclusion criteria. In the third step, we placed mined data in the datasheet based on the critical aspects of our above-mentioned research questions.

To perform a detailed analysis, two researchers independently examined and analyzed the full text of the research articles. Both researchers compared their results and found minor disagreements. Later, both researchers organized meetings to review and settle their disagreements about text extraction. While the third researcher performed data extraction on a random sample, and then he verified and agreed with the results. The analysis work reporting was based on synthesized outcomes. We applied a descriptive synthesis method to explain the results in a manner consistent with our research questions.

Result of SLR process step 4 As a result of the research review protocol step 4, "Analysis including verification", we described our analysed information in the forthcoming sections to answers the above-mentioned research questions.

We depict a remarkable descriptive statistic from our SLR process. Such figures confirms the growing hype around the area of this study. Big data lifecycle is one of the vital study areas among stakeholders. We noticed that digital research libraries, particularly Springer and IEEE, stimulate data lifecycle research works. Our primary studies period distribution of research articles is given in Fig. 3.

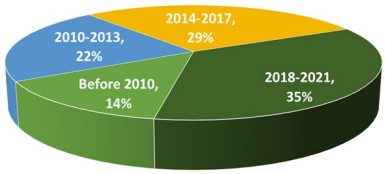

Fig. 3 Temporal distribution of research articles about data lifecycle

Results

The last step of the above-mentioned SLR process describes the research outcomes and presents a foundation for the study. Therefore, we thoroughly reviewed 232 research articles to extract relevant information. In this section, we map our investigation to our research questions. As RQ1 aims to find and organise the literature for our proposed data lifecycle for GBDE, therefore, we present existing related models in this segment. Whereas RQ2 intends to offer a new data lifecycle framework for data-driven governments and its phases, we rigorously explain the approach, considering and building on top of the current state-of-the-art data life cycle models, i.e., an outcome from addressing RQ1, in this section as well.

RQ1: Existing data lifecycle models and their phases

In the literature, we analysed data lifecycles models mentioned in the research articles of our preliminary studies. We selected the 76 data lifecycle models by using the following logical rules: concerned researchers carry out formal research work, offer a sound basis to propose a model, somehow map with industry standards for data management like DM-BOK, relates to areas of government, like public scientific research, open government, and consists of distinctive phases to propose DaLiF. An overview of the data lifecycles and their phases is as below.

Overview of literature data lifecycles We present a few data lifecycle models and their related phases to provide a more detailed understanding. Additionally, we showcase a number of graphical figures indicating findings of the data lifecycles:

A lifecycle for an organization's security data In 2020, A. Shameli-Sendi, presented a lifecycle for an organization's data protection [47]. This data lifecycle consists of the following phases, data creation, edit, display, process, transfer, and store.

IBM data lifecycle model IBM defined a lifecycle model in 2013. This lifecycle model phases include data creation, data use, analysis, data sharing, data update, data protection, archive, store/retain, and dispose [51]. IBM introduced the following additional layers in its data lifecycle model: data masking and archiving [51].

Data lifecycle for OGD In 2019, H. ELMEKKI et al. proposed a data lifecycle that concentrates on Open Government Data (OGD) [98]. This data lifecycle consists of the following phases, data collection, data publication, data transformation, data quality, use, data interoperability, share, user feedback, and data archive.

Research data lifecycle In 2020, R. Raszewski et al. mentioned a data life that is based on UK Data Service Research Data Lifecycle. They aim to find out the extent of data management education in nursing doctoral programs [107]. This data lifecycle

consists of the following phases: plan research, data collection, process and analysis of data, publish data, share data, store data, and use and reuse data.

Abstract data lifecycle model (ADLM) Knud Möller defined Abstract Data Lifecycle Model (ADLM) in 2012. This lifecycle is specifically intended for the semantic web [108]. This data lifecycle consists of the following phases, planning, creation, enrichment, access, store, archive, feedback, and termination phases.

USA (NIST)-big data lifecycle model This data lifecycle model is introduced by the National Institute of Standards and Technology—NIST, Deptt: of Commerce, the USA in 2015. The lifecycle consists of the following phases, collection, preparation, analysis, and action (analytics, visualization, access) [91]. This model concentrates more on the data analytics part compared to data planning, data filtering, and data enrichment.

Research data lifecycle In 2019, M. Jetten et al. described a research data lifecycle [109]. This data lifecycle consists of the following phases, plan, create, process & analysis, use, preserve, and access research data.

OGD lifecycle In 2015, J. Attard et al. mentioned an Open Government Data lifecycle [110]. This data lifecycle consists of the following phases, data creation, data selection, data analysis, data curation, data publishing, data discovery, data exploration, data storage, and data exploitation.

Data Lifecycle for industrial and healthcare applications: In 2020, Kumar Rahul et al. mentioned a data lifecycle for industrial and healthcare applications and focused on managing big data analytics [111]. This data lifecycle consists of the following phases, create, store, analysis, use, share, archive, and destroy.

Web content management lifecycle In 2003, S. McKeever, Dublin Institute of Technology, Ireland, defined Web Content Management (WCM) lifecycle [112]. This lifecycle consists of the following phases, creation, deployment, share, data control, and administration of the contents, storage, archive, use, and workflow.

Research360 Data Lifecycle: Research 360 data lifecycle is defined by A. Ball, University of Bath, UK. This lifecycle focuses on scientific research data, and it is an outcome of the Research360 project [84]. This data lifecycle consists of the following phases, Design, collect & capture, interpret and analyses, manage and preserve, release, and publish, discover, and reuse.

Smart data lifecycle EL Arass et al. define a data lifecycle, identified as smart data lifecycle, in 2018. This lifecycle focuses on how to transform raw and worthless data into Smart Data [60]. This data lifecycle consists of the following phases: Planning, Collection, Integration, Filtering, Enrichment, Analysis, Visualization, Access, Storage, Destruction, Archiving, Quality, and Security.

Hindawi lifecycle Nawsher Khan et al. introduced Hindawi data lifecycle in 2014. This data lifecycle mainly concentrates to applies the tool, technologies, and terminologies of big data [61]. This lifecycle consists of the following phases, collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery.

Data lifecycle for multistage privacy protection in IoT environment In 2019, Soltani Panah et al. introduced a data lifecycle specifically centered on IoT environment [97]. The data lifecycle consists of the following phases, data acquisition, data processing, data storage, use, and data dissemination.

Data lifecycle for information Science: Gislaine P. F. et al. introduced this data lifecycle in 2021. It consists of the following phases: data collection, storage, visualization, and data dispose [113]. This lifecycle focuses on influence mapping of its phases with GDPR principles and how Blockchain technology manages big data in each phase of the lifecycle.

The other existing data lifecycles include: Data Lifecycle for HPC Scientific data perspective [114], Data lifecycle for cloud automation tools [40], Data Lifecycle for Telco networks data management [115], Energy big data lifecycle [116], Data lifecycle for the Tobacco industry [117], Data lifecycle for cloud computing [118];Data lifecycle for cloud data [119], Data lifecycle for IoTs [120], Personal data lifecycle [121], Data lifecycle about the coal mine industry [122], Data lifecycle for smart healthcare [123], Data Lifecycle Model for NSF [94], Data lifecycle cycle for smart cities [13], Storage data lifecycle [124], Research data lifecycle [125], lifecycle for CENS Data [126], data lifecycles for industry [127, 128], a lifecycle for big scholarly data [129], a lifecycle for social and economic data [46], Data lifecycle for manufacturing [130], Research data lifecycle [131], a lifecycle for big healthcare data [23, 132], data lifecycle [133], a lifecycle for environmental research data [134], a lifecycle for big data value creation [26], a lifecycle for big data analytics for psychologists [135], the information pyramid of Reynolds and Busby lifecycle [17], Yuri Demchenko data lifecycle [10], Data lifecycle [58], SCC-data lifecycle [44, 54], Data value cycle [136], Lifecycle in databases [137], Knowledge process-lifecycle [138], CMM for Scientific Data Management process lifecycle [63], Data lifecycle [139], PII lifecycle [17], Data lifecycle [140], Data Lifecycle Process Model [141], data lifecycle for cloud computing security [142], IHRB Data Lifecycle [143], security data lifecycle [144], Knowledge lifecycle for e-learning [145], IRIS Data Lifecycle [62], scientific Data Lifecycle Management (SDLM) Model [146], Big data privacy lifecycle [95], DataONE lifecycle model [15], CIGREF data lifecycle [17], Australian National Data Service (ANDS) Data Sharing Verbs [85], Research Data Lifecycle (UKDA) [53], APDLM - Abstract Personal Data Lifecycle Model [48], Web Data Lifecycle [92, 93], Data Digital Curation Center (DCC) Lifecycle [90, 147], OGD lifecycle [17], IoT Data lifecycle [148], Open Government Data Lifecycle [52], Data Lifecycle for IoTs data [149], Information Lifecycle [88], COSA-DLC [42, 43], USGS data lifecycle [55], and DDI (Data Documentation Initiative) Lifecycle [14, 150]. We describe these data lifecycles in Additional file 2 to meet the journal's page limits.

In Fig. 4, we present some statistics related to our findings. The first figure indicates the distribution of 76 analyzed data lifecycle models concerning the authors' affiliations: industry, academia, and joint authorship. Although most of the data lifecycle model articles were written by researchers, still approximately one quarter of model studies were written by the industry or jointly by academia and industry.

The second figure identifies the source area from where the data lifecycle models research stems. Almost half come from business areas (e.g., health, manufacturing, etc.). One quarter is technology driven (IoTs, smart cities, cloud computing, etc.), a bit less than a quarter from scientific research. Five models depart from the areas of open government and the semantic web.

We also investigated the geographic distribution of the data lifecycle models research. Using the ITU regional classifications [151], the third figure indicates that European

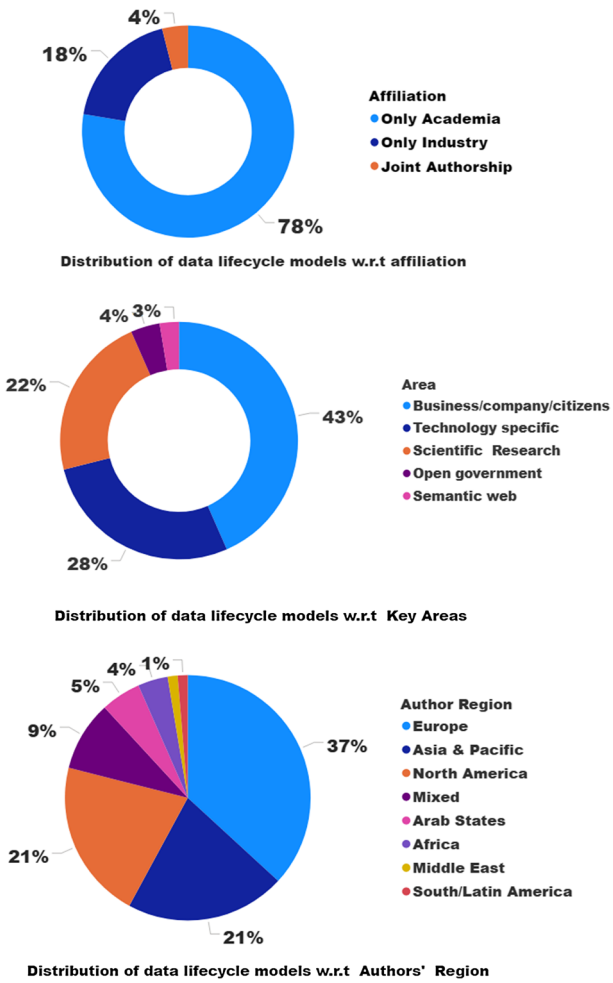

Fig. 4 Graphical figures indicating findings of data lifecycle model studies

researchers contribute more to this body of work (37%), followed by North America and Asia & Pacific regions (21% each).

RQ2: create a big data lifecycle framework for data-driven governments

In this part, we introduce our proposed data lifecycle framework for GBDE. We thoroughly investigated data lifecycles to learn from existing work. Specifically, we wanted to identify commonalities and all the different characteristics introduced by the models to take them into consideration. We explain how we discover and define the phases of DaLiF.

We adopted a comprehensive approach to propose DaLiF. Our approach consists of five steps. In *step 1*, we explain the exclusion of data lifecycles based on logical rules. In the *second step*, we discover and enlist phases from the identified 76 data lifecycles. We already described the selection criteria of the 76 data lifecycles in the preceding section. After a detailed analysis, we ended up with fourteen distinct phases based on certain logical rules. In *step 3*, we group the fourteen phases into mandatory and optional. In the *fourth step*, we apply our analysis to validate the categorization of phases, and we

conclude with six mandatory phases. In the *last step*, we also map phases with DM-BOK functions. We thoroughly describe the above-mentioned steps of our approach to propose the DaLiF as below:

Approach to DaLiF

Our approach to defining the DaLiF consists of the following steps.

Step1: In the first step, we perform a thorough analysis of data lifecycles mentioned in the research articles of our preliminary studies. We excluded 35 data lifecycles based on the following logical rules:

- Data lifecycle does not propose any distinctive phases.
- There is only a description of the data lifecycle and its phases without providing detailed work.

Step2: Enlisting phases for our proposed data lifecycle: In this step, we identify and enlist more than 500 phases, including duplicate phase titles, coming from the analysed 76 data lifecycles. We examined these phases using the following logical rules:

- Grouping phases by title with similar meaning, i.e., synonym detection, or relevant terms used by the various models to propose a phase title that covers the same concept, e.g., for delete, destroy, dispose of, destruction, end of life, we include the phase "end of life".
- Consideration of a single phase that is described in the literature with identical titles, e.g., for visualization [62], visualization [60], visualization [17], we included the phase "visualization".
- Removal of phases that are either too generic, confusing, or just introduced based on a specific research topic by some researchers (e.g., data value, ontology, transformation).
- Removal of phases that were incorrectly labeled, e.g., design, big data.
- Combing phases with similar aims and activities to present a holistic phase with a common objective, e.g., for sharing, publishing, we combine them into a single phase, "sharing/publishing".

The detailed analysis work based on the above led us to fourteen distinct groups/concepts, i.e., phases that we present in Table 1 together with the "source" references from where we find the various phases.

The pictorial representation of the phases matrix with the terms that have been grouped to a more general term appears in Fig. 5. We detail these phases along with their key functions in the next section.

Step 3: Proposing mandatory and optional phases for the data lifecycle: After concluding the above-mentioned fourteen distinct phases in step 2, we further investigate these phases to categorize them into mandatory and optional phases for DaLiF. We adopted the concept of mandatory vs. optional phases of the data lifecycle from the literature. It is a fact that data related to different realms do not all follow the same phases of a lifecycle. Therefore, some phases can be considered mandatory, while other phases are opted

Table 1 Grouping of phases found in the literature

P-(Phase)	Phases as found in the literature	Our Opted phase title
P1	Planning [49, 55, 60, 92, 92, 93, 125], conceptualize [149],plan [15, 17, 53, 55, 58, 84, 107, 109]	Planning
P2	Collection [10, 14, 28, 40, 42–44, 48, 52, 54, 60, 61, 86, 91, 107, 112–114, 127–129, 131, 133, 136, 141, 146, 152–154], collect [15, 53, 58, 134, 143], receive [17], acquisition [13, 117], retrieve [115, 121], produce [140], acquire [55, 63, 90, 97, 147], capture [84, 94, 126, 130, 135, 148], generation [26, 56, 95, 96, 118–120], create [47, 109–111, 116, 124, 149]	Collection
P3	Integration [17, 26, 109, 124, 126, 129, 130, 146, 148, 153, 155], integrate [15, 28, 116, 140, 156], aggregate [115, 120], Integration and filtering [141], Integration and enrichment [134], Integration, filtering and enrichment [10, 17, 60], filtering [121],pre-process [96, 97, 130, 143, 144], preparation [91, 127, 135],classification [61], data description [40, 42], Extraction [120], data selection 351, refinement [92, 93], enhancement [62], cleaning [28], datafication [26, 47, 107, 127, 129, 143]	Preparation
P4	Analysis [10, 14, 15, 17, 26, 53, 55, 60, 63, 91, 98, 107, 110, 111, 116, 121, 126–131, 133–135, 141, 146, 148, 149, 152, 153], rkanalytics [155], data modeling [23, 132], interpret and analysis [84], processing [40, 47, 94, 95, 109, 143], evaluate [140]	Analysis
P5	Visualization [10, 17, 26, 28, 60, 62, 91, 107, 116, 127, 129], presentation [154], knowledge creation [23, 132, 135],data exploration [110]	Visualization
P6	Storage [10, 13, 17, 23, 26, 47, 60, 62, 86, 88, 93, 94, 107, 109, 110, 113–117, 120, 121, 126–133, 135, 141, 143, 146, 149, 152–154], store [53, 58, 66, 90, 97, 111, 119, 124, 134, 142, 147, 148], retention [48], distributed storage [95, 96, 144], storing 161	Storage
P7	Access [49, 53, 58, 60, 85, 88, 90, 92, 93, 109, 127, 143], retrieve [124]	Access
P8	Share [26, 51, 66, 107, 111, 114, 116, 118, 119, 134, 135, 153], transfer [47, 86], sharing [61, 63],dissemination [42–44, 54, 94, 97], delivery [40, 120, 128], data interoperability [98], share [55, 134, 146, 148],publish and share [55, 61], Publish [49, 58, 109, 112, 126, 131, 143, 157], publication [28, 52, 61, 92, 93, 107], release and publish 94	Share/Publish
P9	Use [15, 26, 66, 90, 90, 111, 116, 118, 134, 135, 142], consumption [92] feedback [93] using [53], usage [10, 48, 119, 133, 143, 146, 148, 153], Reuse [28, 61, 84, 88, 107], exploit [85, 110], consume [114], application [130], use and re(use) [13, 49, 147], feedback [52, 140], review [48, 141]	Use, re(use), and feedback
P10	Governance [152], workflow control, and administration [112]	Governance
P11	Quality [13, 15, 60, 98, 133, 134, 141, 143]	Quality
P12	Protection [63, 143], security and protection [60],privacy [23, 97, 121, 154], privacy protection [132]	Protection
P13	Archive [51, 66, 86, 90, 109, 116, 131, 134, 147, 148, 152], archiving 157, 156,117,352,129, data masking [51, 94, 98, 119, 124, 141, 146, 153]	Archive
P14	Destruction [17, 48, 60, 90, 133, 146, 153] end of life [58, 116], destroy [111, 143], Delete [13, 66, 108, 124, 148], dispose [51, 113],termination [49], demise [140], destruct [142]	End of life

as optional [17, 142]. To classify phases as mandatory for the proposed data lifecycle based on the following criterion.

- A phase that appears in most data lifecycles which are found in the preliminary studies.

We classify phases as optional in case a phase does not comply with the above-mentioned criterion. After a thorough analysis, the categorization of the phases based on the above-mentioned criteria is described in Table 2.

We assign a "mandatory" category to phase that has > 70% of appearance in the data lifecycles, and in other cases, we assigned it "optional" category as shown in Table 2. Additionally, we provide detailed information about the data lifecycles and their

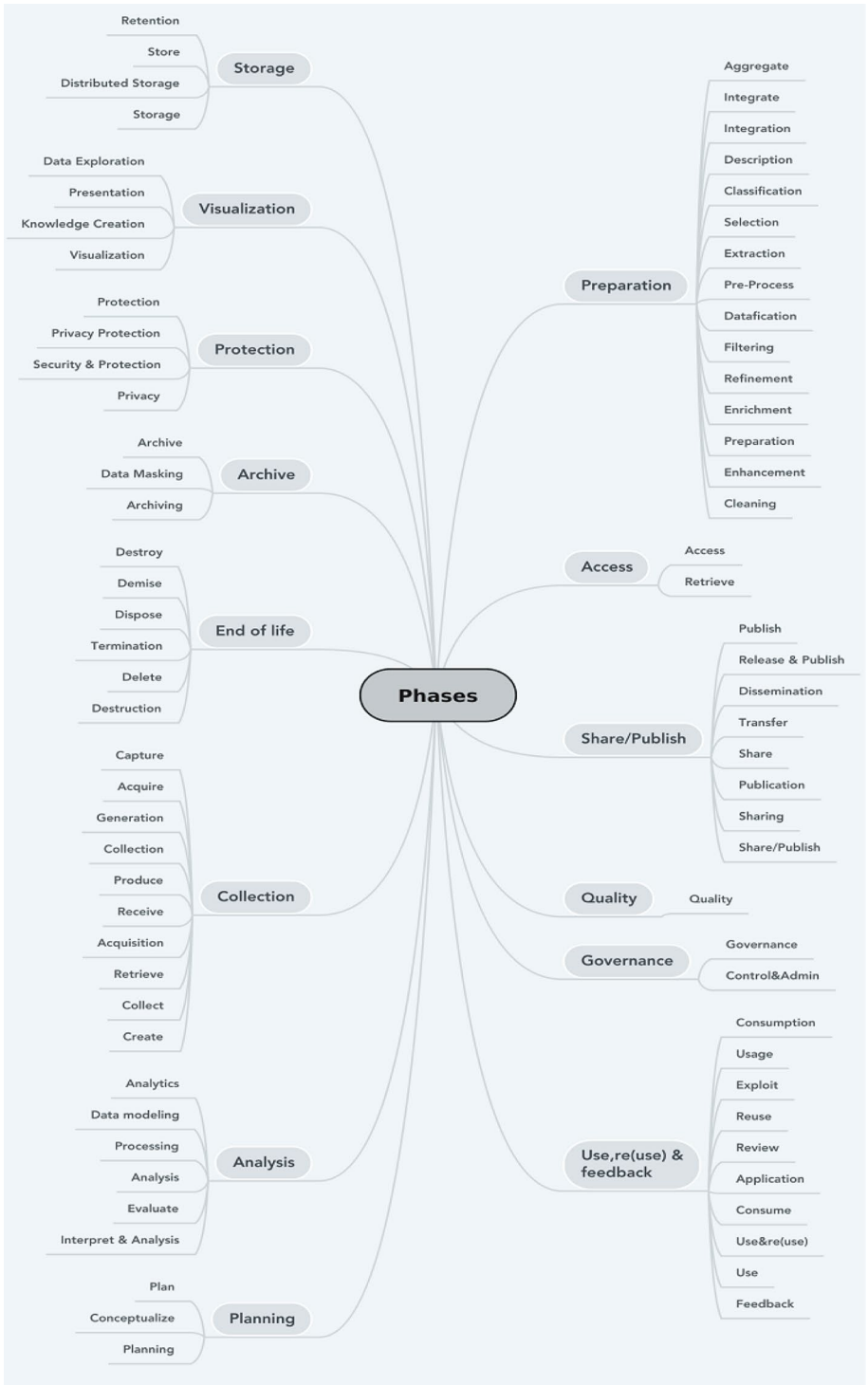

Fig. 5 Graphical view of the proposed grouping of phases

Table 2 Categorization of phases based on their appearance in the data lifecycles

Phase	% of appearance in data lifecycles	Category
Planning	22.37%	Optional
Collection	98.68%	Mandatory
Preparation	73.68%	Mandatory
Analysis	81.58%	Mandatory
Visualization	25.00%	Optional
Storage	82.89%	Mandatory
Access	22.37%	Optional
Share/Publish	78.95%	Mandatory
Use, re(use), and feedback	72.37%	Mandatory
Governance	10.00%	Optional
Quality	19.74%	Optional
Protection	22.37%	Optional
Archive	44.74%	Optional
End of life	25.00%	Optional

Table 3 Five (05) groups of data lifecycles

Data lifecycle models group title
Scientific public research
Semantic web and web contents management
Open government
Business/Company/citizen
Technology focus area (e.g., IoT, cloud computing, smart cities, etc.)

phases in a tabular form, in Additional file 3, wherein the reader can comprehend values in column "%" of appearance in data lifecycles" of Table 2.

Step 4: Validation of the categorization of phases of the data lifecycle: In this step, we apply our own analysis to validate the categorization of phases. In this analysis, we adopted the following criteria.

- A phase that remains relevant in government-related areas data lifecycles

To implement the above-mentioned criteria, first, we identify the main groups of government-related areas data lifecycles. To achieve the said goal, we analyzed 76 data lifecycles. As an outcome of this analysis work, we found the five main groups of government-related areas data lifecycles, as shown in Table 3.

We implement the above-mentioned criteria to validate the categorization of phases of the DaLiF. The phase that remains relevant in government-related these five areas data lifecycle is considered as mandatory, and in other cases, the phase is assigned an "optional" category. The outcome of the above-mentioned validation approach is presented in Table 4.

Step5: Mapping of phases with DM-BOK framework: In this step, we attempt to map our proposed phases with above-mentioned DM-BOK functions [70] as described in Table 5.

Table 4 Validation of the categorization of phases of the data lifecycle

Phase	Remain relevant in 5 government-related areas data lifecycles	Category
Planning	×	Optional
Collection	✓	Mandatory
Preparation	✓	Mandatory
Analysis	✓	Mandatory
Visualization	×	Optional
Storage	✓	Mandatory
Access	×	Optional
Share/Publish	✓	Mandatory
Use, re(use), and feedback	✓	Mandatory
Governance	×	Optional
Quality	×	Optional
Protection	×	Optional
Archive	×	Optional
End of life	×	Optional

Table 5 Mapping of Phases with DM-BOK framework

Phase	DM-BOK concept/function/and related activities
Planning	Data management planning → data governance ; data model and design quality management → data development ; data technology management → data operation management ; define data security policy → data security management; and define data quality business rules → data quality management
Collection	Data Development ; and Create and Maintain Metadata → Metadata management
Preparation	Process Data for Business Intelligence → Data Warehousing and Business Intelligence Management; and Integrate metadata → Metadata management.
Analysis	Predictive Analytics and Data Mining Tools → Implement Business and Intelligence Tools and User Interfaces → Data Warehousing and Business Intelligence Management
Visualization	Advanced Visualization and Discovery Tools → Implement Business ; and Intelligence Tools and User Interfaces → Data Warehousing and Business Intelligence Management
Storage	Implement Data Warehouses and Data Marts → Data Warehousing and Business Intelligence Management
Access	Data implementation → data development; Manage Data Access Views and Permissions → data security management; and Provide Content Access and Retrieval → Document and Content Management
Share/Publish	Query and Reporting Tools → Implement Business Intelligence Tools; User Interfaces → Data Warehousing and Business Intelligence Management; and Distribute and Deliver Metadata → metadata management
Use, re(use), and feedback	Data implementation → data development
Governance	Data governance
Quality	Data quality management
Protection	Data security management
Archive	Archive, retain, and purge data → database support → data operations management
End of life	Data operation management; Retention and Disposition of data/Documents/Records-Document and Content Management; and Data development

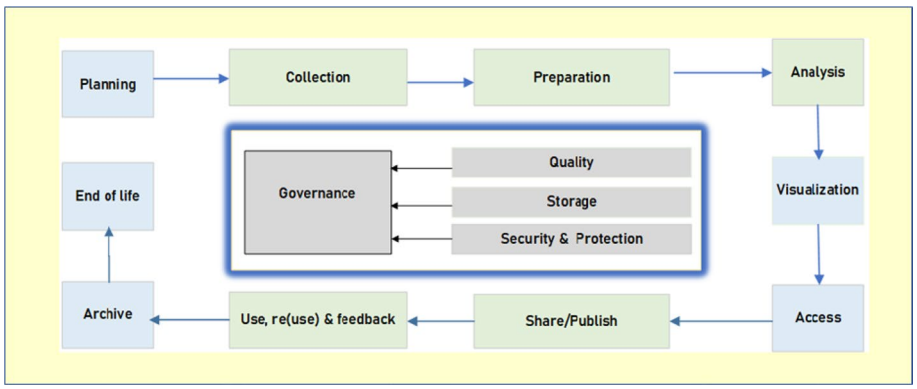

Fig. 6 DaLiF—Data Lifecycle Framework for data-driven governments

Proposed data Lifecycle Framework for data-driven governments

The DaLiF is shown in Fig. 6.

Figure 6 presents the mandatory phases in *green*, whereas phases in *blue* are optional. Moreover, data lifecycle phases, in *gray*, are horizontal phases performed throughout the lifecycle. Moreover, the storage phase is mandatory and horizontal as well. Hence, DaLiF consists of the following fourteen phases, planning, collection, preparation, analysis, visualization, storage, access, share/publish, use, re(use) & feedback, archive, and end of life. In the forthcoming paragraphs, we comprehensively described these phases of DaLiF.

Description of DaLiF phases and their critical functions

In this part, we present the description of the fourteen phases along with the functions of DaLiF. A function indicates an essential activity related to data to be performed by an entity in a phase. We also incorporate the principles of the DM-BOK concept in the respective phases to align our work with the said data management standard.

Planning phase The planning phase illustrates activities to be performed in the medium and long terms during the data lifecycle [6, 53]. This phase consists of formulating a project (e.g., research or business project) to achieve PAs desired goals. Through this phase, it is possible to know the overall objective of the data management, what policies, procedures will be required to treat the data like procedures to collect or generate government data, what data types, sources, and methods are needed to analyze the government data. Additionally, how and where government data is to be stored when such data will be archived or destructed, how it safely will be accessed by the authorized users [46, 48, 49, 55, 149, 158]. The planning phase can help the PAs, including public sector scientific researchers, to save time, boost effective governance, and meet the desired needs about planning for data management in GBDE [6, 53]. The output of this phase is a holistic data management plan [48].

Planning phase key functions The planning phase of DaLiF includes the following key functions:

- Plan for all required resources, including finance and personnel, metadata contents and formats, data storage, data security, and expected outcomes for each phase of the big data lifecycle [15, 53, 55].
- Identification of requiring individuals, descriptions of skills that each of the individuals necessitate to acquire, define roles, and assign roles and responsibilities to the individuals and other public sector stakeholders [53].
- Define a data management plan that is a live document in nature and covers numerous public data aspects like data lifetime, approaches for data quality, data security, and data archive [6, 15, 53].
- Provide a detailed description of data that will be compiled, by whom, and how the data will be managed, made accessible, shared, and reused throughout the lifecycle [6, 15, 53, 62].
- Develop an appropriate plan to select modernized and extensible tools for the data phases, including public data collection [15, 58].
- Plan to prioritize public data that have more possibility for use and be published on the web [93].
- Expert people in the handling of data, records, and content should be fully engaged in planning [70, 71].
- Plan activities that apply quality management techniques to measure, assess, improve, and ensure the fitness of data for use [70].

Collection phase Collection phase consists of set of activities through which data is gathered from different internal and external sources and in different formats i.e. structured, unstructured and semi-structured forms [17, 48, 53, 55, 159, 160]. Big data would have been worthless if it cannot be collected into consistent information [48]. The big data is created or collected from various data resources like social networks, the Internet of things, surveys, census, voting, historical maps, seismology motion sensor outputs, biological records, satellite observations, and commerce statistics [55, 61, 84, 161]. The collection phase defines the moment when the new data or metadata is created in the system [49, 57, 159]. In the Extract, Transform, Load (ETL) common procedure, "Extract" is close to "collect" and this procedure performs a vital role in data collection [162]. In the public sector, during the data collection phase, the PAs should consider the once-only principle to collect data from citizens and businesses and reuse data instead of recollecting [58].

Collection phase key functions The data collection phase of DaLiF includes the following essential functions:

- Collection of raw data from any sources in order to handle big data 'Variety' challenge, and to ensure endpoint input validation to avoid security of data issues [23, 42–44, 57, 58, 91, 132].
- Implement a strategy plan to select modernized extensible tools for public data collection platforms [58].
- Introduce a data protection awareness program while collecting data [23, 70, 132].

- Collection of metadata about information, based on metadata standards, to ensure interoperability across an organization and another future course of action [53, 57, 70].
- Adherence to the once-only principle, in public sector organizations, to collect data once from citizens, and business community [59].
- Manage the ranges of valid and trusted data sources for data collection [42–44, 54].
- Manage the massive amount of data in any format to handle big data volume challenges and search and discover new sources for data collection [42–44, 54].
- Consider specific resources to manage the big data 'Velocity' challenge that refers to the speed rate of data stream generation and the consequent ability to process it well [42–44].

Preparation The preparation phase refers to data integration, filtering, and enrichment [91, 127, 135]. The integration consolidates all these data silos into a single place with a coherent and homogeneous structure [26, 46, 155]. Filtering is focused on the purification of data by filtering the noisy and erroneous data [17, 61]. Data Enrichment refers to the process that appends or otherwise, enhance collected raw data with relevant context obtained from additional sources [10, 63, 143].

The integration enables users to do their queries easily and obtain responses from a single data source [28, 155]. The integration can be considered as a database subarea that provides uniform access to various data sources [163]. Data Lakes (DLs), conceptualize as big data repositories, store raw data, and give functionality for on-demand data integration with the help of metadata descriptions as and when required [162]. However, in DLs, data integration does not take place after the data collection. In the Extract, Transform, Load (ETL) common procedure, "Transform" is close to "integrate" and this procedure performs a critical role in data integration. The integration is based on a set of rules and policies [17]. The data integration is somehow an interim step to achieve a single source of data [17, 28, 164]. In the literature, we noticed that a considerable cost is required for data integration. Such a substantial price is due to data integration across multiple domains, various formats, different vocabularies, metadata of varying quality, and political boundaries [28].

Filtering also allows the data classification in different formats like structured and unstructured formats [61]. The filtered data is further processed through succeeding phases. After due process, policymakers use filtered data to make better decisions within a limited time and with fewer resources [58, 152]. The software development team implements data filters to extract the required data from a vast amount of collected data [62]. The output of the data filtering is a set of categorized, purified, anonymize, less noisy, and error-prone datasets [17, 42, 61, 121].

In enrichment, the normalized, enriched, and simplified data is used through data analysis and mining to generate new information [130, 135]. The enrichment activities are performed on the integrated massive amount of data to limit the selection of data as per certain criteria [28, 61, 62, 90]. The outcome of the data enrichment is a set of refined and mature datasets compared to the original raw data, which can be utilized for either further analysis or for archiving for future inquiries over historical data [42]

Preparation phase key functions The preparation phase of DaLiF includes the following key functions:

- Creation of a homogeneous set of data by consolidating data that is gathered from numerous data sources [6, 60].
- Implement a plan to select modernized and extensible tools for public data integration platforms [17, 58] and tools for data filtering [58, 62].
- Perform scalability about the big data that is high in volume, veracity, velocity and comes from a diversity of sources [155].
- Process big data with the addition of extra measures to achieve a public administration's short-term integration goals [164].
- Do activities, like forming relations among variables of different data sources, adapting units, translating, and building a single database along with all the acquired data, so the government data can be traceable and easier to access for future use [46].
- Consider data privacy protection constraints to avoid revealing private information, like citizen personal and government classified information, in the integrated data [46].
- Identify noise and errors in the collected data and process this public data to remove such issues [61].
- Involve reliable and authorized HR resources in the phase to avoid the leakage of sensitive government data [23, 121].
- Define filtering criteria to be used by the PAs and researchers to filter the public data, including research data as per their needs [61, 62].
- Verification of reliability of the GBD sources as well as of the own data, to manage for any data inconsistencies [40, 42]
- Responsible for carrying out certain fundamental data transformations to optimize the volume of data flowing from the data collection to the quality phases [42].
- Preparation of additional internal and/or external sources data to be merged with existing public data for data valuation [10, 70, 143].
- Extension of existing information by extending missing or incomplete data [61, 63, 130].
- Carefully process data to eliminate unnecessary, misleading, unreliable, and duplicate information [42, 130].
- Establish effective data integration architecture that controls the replication, and the flow of data to ensure data quality and consistency, especially for reference and master data [70].
- Describe source-to-target mappings and data transformation designs for extract-transform-load (ETL) programs and other technology for constant data cleansing and integration [70].
- Implement methods for integrating data from multiple sources and the suitable metadata to make sure meaningful integration of the data [70].

Analysis The analysis phase is the most common phase of all data lifecycles models. The analysis phase enables an organization to handle ample information that can affect the business [61]. This phase is responsible for developing all data analysis and data analytics

to extract knowledge and discover new insights [34, 42, 165]. This phase is like a human brain, i.e., processes the information for the next appropriate required actions by human beings [143, 155]. In this phase, different big data analytical tools are utilized to analyze data. The data analysis tools help the policymakers to analyze much and complex data to understand what is happening (descriptive analytics), to understand the reasons that something is happening (causality) to run what-if scenarios, and to do forecasts. The policymakers use forecasts in their decision making [143, 155]. Modern technologies, state-of-the-art data infrastructure, and highly skilled people are critical to extract relevant insights from big data in the analysis phase. Examples of modern technologies include machine learning, deep learning, artificial intelligence and natural language processing. Relevant people's skills include data analytics, data mining, computing, statistics, etc. [152]. The analysis phase includes analysis of unstructured data [91]. The output of the data analysis phase includes knowledge, the discovery of new insights, new data, interpretations and/or new datasets [42, 54, 61, 165].

Analysis phase key functions The data analysis phase of DaLiF includes the following key functions.

- Data sources selection like identification of descriptions, data sources location, file types, and data provenance [58, 91].
- Perform analysis of data to extract knowledge and discover new insights, and then decision-makers use this knowledge intelligently to generate value for public organizations [17, 42–44, 54, 61].
- Consider innovative data analysis strategies like schema on read to manage public data through mode tools [91].
- Select appropriate data analysis tools and techniques, like data mining algorithms, cluster analysis, correlation analysis, statistical analysis, and regression analysis, to analyse public organizational data [58, 61].
- Set-up a data scientists' group (actors) with sound expertise in data analytics to perform analysis of various types of public data, particularly unstructured data, and describes a set of actions to be completed by the group [61, 62, 91, 152].
- Discover business processes that can be enhanced through big data technology, perform analysis of existing issues in each business process, and perform re-engineering of each business process using big data technology [62].
- Define the required types of big data analysis that include descriptive, predictive, prescriptive, and diagnostic [34, 62, 165].
- In the analysis phase, prepare and publish outputs in machine-readable formats [53, 55].
- Extraction of value from big data through its extensive use and offers a natural interface with the data users [35, 42–44, 54].

Visualization The visualization phase deals with the presentation and visualization of the outcomes, as well as explanation of the meaning of the discovered information [62, 91]. The visualization phase has the highest value for the data consumer in the information value chain, and this phase also boosts the interaction of data analytics and the organizations [91, 155]. We also noted the following categorization of data visualizations,

exploratory data, explicatory, and explanatory visualization. The first category emphasizes better data understanding, particularly in a huge amount of repurposed data. This is because of the volume of the datasets that need new methods. Examples of exploratory data visualization include browsing, boundary conditions, and outlier detection. The second category focuses on analytical results. Example of explicatory visualization includes confirmation, interpreting analytical results, and near real-time presentation of analytics. The last category is about 'telling the story' and a simple way of presenting results to the layman to ingest easily. Examples of explanatory visualization include business intelligence, reports, and summarization [91]. This phase results can be offered in various forms like dashboard, oral presentation, user interactions, alerts, reports [62].

Visualization phase key functions The visualization phase of DaLiF includes the following key functions.

- Visualize public data so that less tech-savvy decision-makers can understand and use results for effective decision making [64, 91].
- Implement a plan to select modernized and extensible tools for the data visualization, like pipes in 'R' computer programming language and geoms (Cleveland dot plots, box plots, and jittered graphs) [58, 64].
- Encrypt the resulting information and knowledge and adopt an access control strategy to avoid privacy threats [23, 132].
- Adopt appropriate mechanisms for reporting and analysing the data, including online and web-based reporting, BI scorecards, ad-hoc querying, OLAP, and portal [70].

Storage The objective of the storage phase is to save data securely throughout the life cycle. Data storage is an essential process of big data analytics in real-world applications [65, 166]. We noticed in the literature that the storage phase is considered in all data lifecycles models. There is a demand for stable and usually web-accessible storage [90, 144, 147]. Data Lakes ingest raw data in its original format from various data sources, meet their role as storage repositories, and allow users to query and explore them to extract knowledge [167]. In the Extract, Transform, Load (ETL) standard procedure, "Load" is close to "store," and this procedure performs a critical role in data storage [162]. The activities of data lifecycle phases, like data access, publish, data sharing, data use, and re(use), would be executed once the data is stored in a place [85]]. However, big data storage is also a complex, costly, and challenging data lifecycle phase. In the public sector, government entities usually setup base registries to store GBD of particular importance i.e. master data. A base registry is a reliable and authentic source of information about people, health, vehicle, crime, and businesses [58]. The base registries support PAs to eliminate data silos and maximize the data's re-use across the public sector entities easily and inexpensively [15, 58]. In the storage phase, different modern tools and technologies are required to store big data like NoSQL, NewSQL, Big Data Query Platforms, Hadoop Distributed File System (HDFS), and cloud storage technologies [50, 95, 96, 168, 169]. Moreover, several NoSQL technologies, like HBase, MongoDB, Cassandra, CouchDB, DynamoDB, Riak, Redis, Neo4J. These technologies store data streams in a real-time fashion into a NoSQL database [127].

Storage phase key functions The storage phase of DaLiF includes the following essential functions.

- Identify public data to be stored, specify a data repository or data center where the shared data will be stored [15, 58].
- Develop and implement an appropriate, short & long- term storage plan to store data in GBDE [53].
- If relevant, ask permission from citizens and businesses to store data of their property [66].
- Store data in an appropriate location (in-house data center or private cloud environment) in a secure, scalable, accessible, and reliable manner [65, 90, 147].
- Compliance with industry standards, a) to store GBD along with improved data structures, appropriate cloud data security, and backing fault tolerance; and b) to improve data storage systems performance in terms of capacity and speed [65, 75, 142, 166].
- Implement a plan for the selection of modernized and extensible tools for data storage along with a balanced approach for data availability and scalability [58, 64, 95].
- Establish base registers to store public data at national and cross-border levels [58].
- Perform continuous work on data storage with improved data structures and backing fault tolerance [65, 85].
- Adopting approaches based on encryption techniques ensures privacy protection in the data storage phase [95, 96].
- Implement a document and content management system that offer electronic documents and electronic images of paper documents storage, versioning, security, metadata management, content indexing, and retrieval capabilities [70, 71].

Access The data access phase focuses on ways of communication between the data provider and data consumer in the big data ecosystem [60]. Through this phase, we decide and document which user [60, 147] or re-user [90, 147] is accessing which data and with what mechanisms [58]. Public sector organizations offer multiple channels for data access [94].

Access phase key functions The access phase of DaLiF includes the following essential functions.

- Ensure the access of public data to users and reusers on a day-to-day basis as per agreed and signed an agreement [60, 90, 147, 149].
- Define data access controls, and data authentication methods [58, 90, 117].
- Establish data access models like cloud, intranet, and virtual desktop models that help determine the hosts' identity, authority, clarify the operation authority, and identify, authenticate remote users, and ensure secure communication, respectively [117].
- Ensure that data that is openly accessible to all users may not by any means contain classified privacy information to avoid personal data privacy threats [109].

- Ensure that limitations on access are conveyed and admired [17, 90, 147].
- Allow government data exchange platforms, like Belgium platform 'MAGDA', to further facilitate data access and exchange of data among public bodies [58].
- The mission-critical data that need to be accessed frequently by the analytical tasks should be stored to offer fast retrieval and updates. While less urgently accessed data can be stored in a database, on disk, or in data files [120].
- Implement dynamic and scalable access control like Authenticator-based data integrity verification techniques [23, 132].
- Enable effective and efficient access and use of data and information in unstructured formats [70].
- PAs should allow access to documents/records in accordance with related policies, standards, and legal requirements [70].

Use, re(use) and feedback In this phase, we combine two key concepts, use, re(use), and feedback. The 'use & re(use)' concept is about the use and re(use) of data by the data consumers [118, 142, 161, 170] and focuses on discovering new and valuable information from existing public datasets by different stakeholders [58]. While in case of second concept feedback, data users exploit the open government data and provide their feedback [49, 98, 98]; such feedback is in the form of user reactions, comments, and suggestions that usually identify improvements and corrections in the published data or metadata [49, 52, 98]. Moreover, re(use) is a process, not a single action, and it includes different activities like acquiring datasets from various public or private data sources to compare to recently collected data, returning to one's own data for later comparisons, surveying available datasets as background research for a new project, or steering reanalysis of one or more datasets to address new research questions [171]. The examples of data consumers include citizens, individuals [67, 118], businesses, researchers, and employees of other government agencies [168]. In the use, re(use), and feedback phase, PA is not the main actor, but the client as the PA can still use and re(use) the public data. App developers create new and valuable information by pulling the non-classified government datasets together and mashing up with other private data to build high-value Apps [28, 142]. The governments are also being working to open data without personal attributes so that businesses and the community use and re(use) such data for innovation, accomplish their day-to-day tasks, and gain commercial benefits from this data [58, 170]. There are a variety of open datasets that are usually used for several objectives by various users. The data publishers usually ensure that their data, incredibly private data, is accessible to designated data use and re(use) [90, 147]. There are different motivations of the data users and re-users like community welfare, business growth, and earn money [11, 28, 172]. The European Commission advised the European Member States to formulate a holistic big data strategy, including publishing open data and promote the use and re(use) of such data. Moreover, the Commission offers special proposals to them to achieve better data use and re(use) within a State and cross-border as well [66, 142]. The other government entities may use and re(use) GBD as a tool to improve and optimize the internal processes of the public administration and make evidence-based decision-making to improve their public services for the public [58, 66]. This phase's output is

a set of manipulated data values [48, 147]. Data feedback is a way to obtain a consensus among stakeholders, including the community. The data providers examine the user feedback about data and again publish modified data after incorporating the data users' feedback [98, 173]. The PAs can gather a vast amount of all stakeholders' viewpoints, as evidence-based information, on public data [58, 120].

Use, re(use), and feedback phase key functions The use, re(use), and feedback phase of DaLiF includes the following key functions.

- The data provider may provide data to the data consumers to use, re(use), and offer feedback about data along with an appropriate mechanism that enables an individual to manage and control their digital record of information [67, 142, 171].
- Ask for permission to citizens, and businesses, i.e., owners of private data, to use and re(use) data of their property consistent with the objectives of information collection [66, 118].
- Outreach all stakeholders so that everyone has an equal chance to provide feedback [49, 66, 153].
- Allocate enough time to the stakeholders and actively listen to them to provide their feedback [58, 98].
- The data provider implements data usage policy, relevant national and international regulations about data use and re(use), and creates awareness amongst about the said policy within data consumers to avoid individual data misuse [58, 66, 67].
- Adoption of consistent and uniform approach(es) and shared (interoperability) platforms to help the safe, transparent, and controlled use and re(use) of data across public organizations. These approaches and platforms also help to discover what data is available and facilitate its use and re(use), preventing duplication of effort across public organizations [58, 142].
- Interact in a more civilized and less bureaucratic manner with the stakeholders to get fruitful and enough feedback from them [52, 66, 153].
- Implementation of base registries, single authoritative sources of data, to enable data use and re(use), and decrease the requirement for citizens and businesses to give the same information to public organizations again and again [58, 118, 152].
- Implementation of the plan for the selection of modern tools and technologies, including API-based technologies to promote data use and re(use) with data harmonization and consistency [58, 66, 161].
- Develop IT systems, connectivity infrastructures, and platforms to proceed towards a country that functions as a unit and increase the use and re(use) of GBD for the decision making [58, 66, 119].
- Ensure the use of technological solutions and social media so that data providers, like PAs, can create informally and efficient ways of communication with data users' including citizens [49, 52, 98, 120].
- Establish possible collaboration with the citizens to express their interest and offer feedback about data published by the government [98, 133].
- Facilitate easy and inexpensive reuse of data across the organisations, preventing, wherever possible, redundant and inconsistent data [70].

Share/publish In this phase, we combine publishing and sharing concepts of traditional peer-reviewed publication with the distribution of data and information through (government) web portals, social media, data catalogs, eGovernment information systems, and other venues [55, 61, 128]. Data and its resources are collected, prepared, and analysed for sharing and publishing to benefit the stakeholders. The examples of such stakeholders include governments, businesses, citizens, researchers, scientific partners, and federal agencies [9, 58, 61]. The data provider shares data with the above-mentioned stakeholders, as per defined ethical and legal specifications [58, 67, 128]. In the government sector, organizations have data related to tax revenue, health, education, economics, transport, etc. The government organizations share data with the rest of the government entities. Data sharing is helpful to achieve greater efficiency in the use and re(use) of data by the government [35, 142, 152]. It is a key to transparency and economic growth [174]. The fundamental idea of linked data is to use the World Wide Webs global architecture to share structured data worldwide [26]. In this phase, the data publish concept emphasizes what data can and should be made public and how data needs to be published with appropriate security measures and integrity [58, 92]. PAs determine which data is to be issued for other government departments and which information is to be disseminated openly to the public [58]. However, PAs do not publish various data sets due to certain data traits, like data containing personal or sensitive information [175]. PAs intend to publish government data for all to promote transparency, accountability, value creation, i.e., better governance, and to enhance the quality of life of the citizens [67, 79, 175]. The data publish phase is highly essential for the open government domain. This phase's output is publishing non-classified data [92, 93].

Share/publish phase functions The data share/publish phase of DaLiF includes the following key functions:

- Implement a plan for the selection of modern tools and technologies, including API-based technologies, to promote data sharing/publishing with stakeholders safely and effectively [58, 67, 128].
- Identification of non-classified public data to be shared or published [58, 92].
- Sign off data sharing agreements between governments and other stakeholders that emphasize the legitimate basis and logic behind why public data is being shared [58, 66, 115].
- Ensure to take appropriate measures that enable individuals to control whom to share data and how much the owner is eager to share [67, 97, 114, 142].
- Data providers focus on maintaining a balance between data availability and data redundancy when publishing data through various formats [79, 93].
- Consider data sharing granularity and data transmission in addition to authorization of data while sharing private data. As sharing granularity refers to conformity to sharing policy and data transmission indicates the isolation of sensitive information from the original data. This function makes the data is not related to the data owners [118].
- Follow open data publishing guidelines and principles as mentioned in [176] and [177] to publish open data [93, 175].

- PAs should keep balance to allocate powers to a different group of stakeholders (Government bodies, NGOs, Regulators, Data Brokers versus data subjects, entrepreneurs, archivists, data, data collectors) in driving the design, framing, and implementation of data sharing policies and practices [174].
- Implement web standards in data formats, like HTML, XML, RDF, CSV, and web protocols, like HTTP, FTP, and SOAP to publish data on web [70, 92, 175].

Archiving phase Archiving is a process to anchors a chunk of data within a system through cataloging, indexing, or a related action [49]. Archiving is for obsolete data, keeping for records in case access is needed, however, at a low storage cost. While data storage is for active information, available for day-to-day activities, but at a high storage cost [61]. Data archiving is one of the prime phases of a big data lifecycle [10]. It is pertinent to mention that effective data lifecycle management includes the intelligence not only to archive data; however to archive the data based on specific parameters or business rules. An example of such parameters consists of the data's age or the last date of their use [51]. In a cloud computing environment, archiving is a technique to shift less frequently used data to another place in cloud s for an extended period [88, 142]. Data Archiving can also help storage administrators to develop a tiered and automated storage strategy to archive static data in a warehouse. Through this strategy, data warehousing specialists can improve overall data warehouse performance [51]. Some researchers describe the data life cycle by the data access frequency [142]. As the moment goes on, the data access rate gradually declines, and ultimately such data goes to an archived state. Additionally, in this phase, the following three main operations are required, encryption techniques, long-distance storage, and data retrieval mechanism. These operations permit the least used data to be shifted to separate storage devices for long-term storage. The archiving and storage devices are thus separated [61, 88]. Some countries have special national archival legislation to archive the government record/data for reference and future use by PAs.

Archiving phase key functions The archiving phase of DaLiF include the following es research work provides a holistic view of 76 datsential functions:

- Data, including personal data, should be archived with strict security measures to avoid such data leakage [58, 142].
- Implement a formal agreed plan to archive data to ensure data availability and data re-use [55].
- Use of appropriate archival standards like General International Standard Archival Description (ISAD-G) for various purposes, including hierarchical data description [90, 147].
- Implement a plan to select modernized and extensible data archiving tools [49, 66].
- Adopt an appropriate archive method to ensure that such data is accessible to the data scientists for data analytics reasons as and when require [51].
- Data resources, data infrastructure, and data management should be forecast to deliver continuity and archive data for as long as required [66].
- Use of appropriate anonymization techniques like generalization and suppression to protect data privacy during this phase [95, 96].

End of life phase: In this phase, duplicated data, no longer required data, and useless data is removed from the system [50, 58, 88, 111]. Data must be considered in terms of end of the usefulness of data or end of life [58, 116]. In the cloud environment, to maximize resource usage, the storage location of data is often moved. As data is moved, the original location is also destroyed [117, 142]. Such titles include deleting, terminate, destroy, and dispose of. Data-driven Public Administrations always make decisions regarding the end of life of data based on their data strategy [58]. This phase's output is a set of destructed data values [48, 88].

End of life phase key functions The end of life phase of DaLiF includes the following key functions:

- Useless, inactive, and data that has attained the end of its lifespan may be destroyed as per rules/regulations [58, 88, 116].
- Implement a plan to adopt appropriate methods for the data end of life [49, 66].
- Data centers, including government data centers, should offer suitable data end-of-life functions like disk replication and demagnetization to their clients to avoid sensitive public data leakage [88].
- Ensure that unnecessary data is permanently removed and cannot be restored from the storage medium to avoid inadvertently disclosing sensitive information [118].
- Ensure that data in the cloud is removed, through appropriate means, according to the owner's mind, to guarantee the information not be disclosed or recovered [119, 142].
- Ensure wiping of unwanted data on partitions and hard disks [118, 142].

Data quality phase The quality phase focuses on maintaining data quality during the whole data lifecycle, i.e., data collection, data integration, data analysis, data publishing, and data share phases [42, 62]. A primary data quality management principle is that manage data as a core organisational asset [70]. Data quality is one of the prime issues related to the value of data for the business [54, 178]. DMBOK highlighted the following dimensions of data quality, accuracy, completeness, consistency, currency, precision, privacy, reasonableness, referential integrity, timeliness, uniqueness, and validity [36, 70, 71]. The data-driven public administrations can offer better services and policies through improving data quality [43, 66, 152, 179]. When we have well-defined quality requirements, then implement controls to measure the data quality's satisfaction is more feasible. Examples of such quality requirements include margins of errors and the requisite level of precision [17, 60]. The quality of the big data is essential for their consumption. Data may be precise, timely, and in accordance with actuality [66, 180]. Base registries (public sector master data) are needed for valuable and highly reusable data [44, 58]. The United Nations also described a set of actions for the computer scientists to ensure quality during data input and output results to limit the risks in various factors like complexity, speed, accuracy, validity, and clarity [62].

Quality phase key functions The data quality phase of DaLiF includes the following essential functions:

- Certify that public data, information, and metadata are of high quality by engaging data quality and metadata experts [60, 66, 70].
- Establish quality criteria and quality processes that consider generation, storage and processing [62, 66].
- Implement data quality management policies, international standards, procedures, and guidelines to cross-check the data quality level to discard the data with low quality, improve the data quality, etc. Such implementation ensures high quality, consistency, the integrity of public data, and help to handle the 'Veracity' challenge [17, 42–44, 54, 66, 152].
- Monitor the data quality flows, in case of failures, then proceed as per the data quality management policies [42–44, 54].
- Apply conformance checks to data quality business rules, like attribute domain constraints, format constraints, and standardisation constraints, at each phase of the data lifecycle to avoid low-quality data, like missing attribute values, schema, and data format differences [43, 181].
- Create and promote data quality awareness within an organisation [70].
- Make explicit data quality attributes like accuracy, integrity, completeness, and timeliness to help policymakers determine whether data is reliable for the decision-making process [13, 180, 181].
- Being business process owners, PAs should agree to and abide by the data quality SLAs [70, 73].

Protection phase This phase focuses on data protection in terms of data integrity, security, access control, and privacy [17, 61, 182]. The phase is being considered throughout the data lifecycle, i.e., from planning, collection to archive, and destruction phase to maintain data security and privacy protection against any accidental or malicious compromises to the GBDE [58, 69, 97, 144, 183]. Due to the quantity, variety, and sensitivity of the big data and its management through heterogeneous based technological solutions, data security and privacy protection become crucial. A holistic methodological approach is required based on data protection standards and common practices to deal with these issues [69, 118]. Adequate data security and privacy protection management establish governance mechanisms that are easy enough to abide by on a daily operational basis [70]. Classified data must be secured and protected from unauthorized users through various data masking techniques to protect data from unauthorized access [17, 51]. There is an in-balance between privacy and the risk of malicious data exploitation [23, 58, 61, 183]. The processing of personal data in Europe is subject to the General Data Protection Regulation (GDPR) and the Data Protection Act of 2018. Such legislation ensures the privacy of citizens and the secrecy of data and information gave by businesses [58, 66, 118]. This phase's output is secured and protected data [61, 144].

Protection phase key functions The data protection phase of DaLiF includes the following essential functions.

- Government organizations should process data in a way that certifies the protection of personal data against unauthorized or unlawful data handling [58, 66, 96, 118].

- Implement privacy standards, introduced by ITU, CSA, ISO, etc., privacy policy, techniques, and security solutions to protect data, including personal data, to avoid data threats. Whereas such solutions and methods will be based on various security patterns like encryption, authentication, anonymization, and role-based access control [23, 69, 70, 118, 119, 132].
- The use of unique identifiers to manage users' digital identities, their relationship to a real-world identity, and access to systems, data, and information are essential for data protection [58, 118].
- Data and information must be protected as prescribed by both regional (like EU) and national (like Italy) legal codes and data protection policies with suitable levels of data protection, security, confidentiality, privacy, integrity, and availability [183].
- PAs should also allocate sufficient funding, create awareness amongst the people, impart requisite training for the staff, and engaged technical experts to protect GBD [96, 118].
- Minimize the risk of privacy violation during the data collection/generation by appropriate means like restricting access or falsifying data [95, 96].
- Ensure privacy protection in the cloud environment by the strict separation of sensitive data from non-sensitive data [118].
- The PAs should take security and data protection processes to identify and protect citizen and business data; for example, privacy-by-default and privacy-by-design will be adopted. [58, 66].
- Ensure double encryption data system using an appropriate encryption algorithm, like AES and RSA, to avoid data mining-based security attacks [23, 132].
- PAs should form arrangements to identify and utilize security requirements applicable to the receipt, processing, physical storage, and output of data and classified messages [70].
- Execute effective data security policies and procedures to assure that the right people can use and update data in the right way [70, 73].
- Collaborate with stakeholders (e.g., IT security administrators, data stewards, internal and external audit teams, and the legal experts) for defining data security requirements and data protection policy [70].
- Adopt data protection tactics at data consumers, systems, and data providers levels to ensure data protection from unauthorized entities, systems, and un-trusted data providers, respectively. Examples of such tactics include personal data stores, software/hardware-based virtualization, data encryption [182].
- PAs should promote the concept of decentralization and private-by-design IoT through Blockchain technology in IoT-based Information Management Systems to ensure data security and privacy protection [184].

Governance phase Data governance phase refers to a plan to guarantee that high-quality and protected data exists and exploited throughout the complete data lifecycle [58, 185]. It determines the policies and procedures to safeguard preemptive and effective management of data assets [70]. Data governance interacts with and influences each of the surrounding phases and guides how activities in other phases are performed [70–72]. Data governance helps the PAs manage data in the public sector organizations as it also

implies the allocation of decision-making rights and associated functions in such management [68, 186]. The data governance phase helps PAs protect public sector organizations' data assets to assure generally understandable, accurate, complete, reliable, protected, anonymous, and discoverable government big data. It is also assisting in systemizing these organizations by linking business processes with data in GBDE [185, 187]. The data governance phase includes consistent management and helps public administrations set data rules/policies, provide insights, wisdom & judgment, and promote accountability [61, 152]. The estimation of the quality of data is recognized to be crucial for data governance. Data governance is one of the central pillars of the data-driven government. Through excellent data governance, public administrations will guarantee that their data are precise, reliable, comprehensive, available, and secure [58, 66, 186].

Governance phase key functions The data governance phase of DaLiF includes the following essential functions.

- Utilize standards, guidelines, tools, policies, laws, procedures, roles, and responsibilities for public data governance to ensure data utility by the data consumers [58, 68].
- Establish a formal system of accountability for effective data governance [58, 152, 187].
- Apply machine learning and AI algorithms to improve data governance [66, 187].
- Create a collaborative environment within the stakeholders, including users, so that public administration will get proposals from them on improving the data life cycle, particularly in case of agility in work scope [62, 186].
- Focus on the following aspects, data quality, data security, and privacy protection to tackle data governance-related issues in cloud computing environment for better visibility, data quality, and protection control [186, 187].
- Promote the use of use machine learning and AI to reframe data governance to address related business requirements in a way that motivates data producers and consumers to work together [68, 185].
- Constitute a Governance Board or a Committee in the organization to oversee and drive data governance across the public services [58, 66, 186].
- PAs must take an organizational perspective to ensure the quality, security protection, and effective use of government data [70, 72].

As an outcome of the research review protocol's step 5, "results", we mentioned our comprehensive research results and proposed DaLiF in the preceding sub-sections of this segment.

Research implication

Given the above-mentioned outcomes, we offered the following research implications for the scholars and practitioners:

Benefits to the research community

- This research work provides a holistic view of 76 data lifecycles and their phases to the research community.

- We proposed DaLiF based on various literature data lifecycles to help the research community in studies related to data management in GBDE.

Benefit(s) to entrepreneurs

- This study may offer insights to entrepreneurs to gauge new business ideas and innovation in developing government big data management solutions and services.

Benefits to the practitioners

- DaLiF could help practitioners to plan and handle the complexity of data management, identify essential activities and maintain appropriate data quality throughout the data lifecycle.
- The detailed overview of DaLiF, including our proposed functions for each phase of the lifecycle, as an information tool, could be workable for the public sector organizations to develop or modify their strategic measures to manage GBD efficiently.

Benefits for the Public Administrations—PAs

- DaLiF could support PAs to gather, classify, refine and analyze data to extract knowledge, find new insights, generate value for the public sector organizations and promote data-driven decision making.

Conclusion and future work

In this work, we propose DaLiF as a big data lifecycle model. The key characteristics of DaLiF include:

- The DaLiF is based on the analysis of 76 data lifecycle models presented during the last 25 years.
- It is not mandatory to follow all the proposed phases as the optional ones can be selected based on the specific needs.
- The model remains relevant and applicable in various government fields such as health, agriculture, education, manufacturing while it covers different types of data, including open government data, business data, scientific and research data, citizen (i.e., personal) data, etc.

Limitation and future work

We summarise here two limitations of our work. Due to limited existing research, we could not find a good number of research articles explicitly on the data lifecycle frameworks for the data-driven government. Therefore, we borrowed concepts from existing literature in "neighboring" areas, i.e., scientific research, Semantic web & web contents management, open government, cloud computing, IoT, etc. Moreover, we examined IEEE, ACM, ScienceDirect, and Springer digital research libraries as we considered them more relevant to our research; nevertheless, other digital libraries may also contain some relevant research articles.

We intend to continue our work in the area. This research article primarily contributes to the theoretical realm and needs practical validation to bridge the gap between academic rigor and industrial applicability. Secondly, in our previous published study on GBDE [81], we presented a classification model for data actors. We have the interest to establish a linkage between data actors and the proposed here DaLiF phases. Lastly, we intend to conduct a detailed survey of technological tools for each phase of DaLiF.

Abbreviations

GBDE: Big Data Ecosystem (GBDE); DMBOK: Data Management-Body of Knowledge; PAs: Public Administrations; DaliF: Data lifecycle framework for data-driven governments; BDE: Big Data Ecosystem; SLR: Systematic Literature Review; IHU: International Hellenic University; PEO: Population/Problem, Exposure, Outcome; FINER: Feasible, Interesting, Novel, Ethical, Relevant; ITU: International Telecommunication Union.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-021-00481-3.

Additional file 1. It is simple file which contains additional text about formulated search strings and we used these strings in digital research libraries search queries about BDE and existing data lifecycles.

Additional file 2. It is simple file which contains additional text about description of existing data lifecycles.

Additional file 3. This is simple file which contains a multi-page table. In this table we provided detailed information about the data lifecycles and their phases.

Acknowledgements

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, and by 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

Authors' contributions

SIHS worked on the conception and design of the paper. He also drafted the paper. SIHS and IM conducted the review and data collection. Both researchers comprehensively analyzed the full text of the studies. While VP validated and verified this research work outcome. He also refined the concepts and proofread the paper as well. All authors read and approved the final manuscript.

Authors' information

Syed Iftikhar Hussain Shah is researcher and Ph.D candidate in IHU, Greece. He has more than 15 years plus professional experience in the area of eGovernment, data science, digital public policy.

Vassilios Peristeras is working in Council of the European Union, General Secretariat, Brussels, Belgium, and he is also Asst. Prof. in IHU, Greece. He published 100 plus research articles.

Ioannis Magnisalis is working as a consultant at DG Informatics, European Commission and he is also teaching web technologies, data science, and e-commerce in IHU.

Funding

Open access funding will be provided by The European Union-funded Projects: Co-Inform, Horizon 2020—the Framework Programme for Research and Innovation (2014-2020); Grant Agreement 770302 and Digital Europe for All (DE4A), Horizon 2020—the Framework Programme for Research and Innovation (2014-2020); Grant Agreement: 870635.

Data availability statement

Not applicable

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details
[1]School of Science and Technology, International Hellenic University, Thessaloniki, Greece. [2] Council of the European Union, General Secretariat, Brussels, Belgium. [3] DG Informatics, European Commission, Brussels, Belgium.

Received: 21 March 2021 Accepted: 2 June 2021
Published online: 14 June 2021

References

1. Becker MJ. The consumer data revolution: the reshaping of industry competition and a new perspective on privacy. J Direct Data Dig Market Pract. 2014;15(3):213–8. https://doi.org/10.1057/dddmp.2014.3.
2. Fabijan A, Dmitriev P, Olsson HH, Bosch J. The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale. Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering, ICSE 2017, 2017, p 770–780. https://doi.org/10.1109/ICSE.2017.76.
3. Fahy R, Van Hoboken J, Van Eijk N. Data Privacy, Transparency and the Data-Driven Transformation of Games to Services. In: 2018 IEEE games, entertainment, media conference (GEM). IEEE, Galway, Ireland 2018, pp. 1–9. https://doi.org/10.1109/GEM.2018.8516441.
4. Aftab U, Siddiqui GF. Big data augmentation with data warehouse: a survey. In: Proceedings—2018 IEEE international conference on big data, big data 2018, 2019, p 2785–94. https://doi.org/10.1109/BigData.2018.8622206.
5. Pathak AR, Pandey M, Rautaray S. Construing the big data based on taxonomy, analytics and approaches. Iran J Comput Sci. 2018;1(4):237–59. https://doi.org/10.1007/s42044-018-0024-3.
6. Allard S. DataONE: Facilitating eScience through Collaboration. J eSci Librarian. 2012;1(1):4–17. https://doi.org/10.7191/jeslib.2012.1004.
7. Mazumdar S, Seybold D, Kritikos K, Verginadis Y. A survey on data storage and placement methodologies for Cloud-Big Data ecosystem. J Big Data. 2019;6(1):15. https://doi.org/10.1186/s40537-019-0178-3.
8. et al S. Government big data ecosystems a systematic literature review. In: International conference on digital information management, Italy, 2020, p 1–14.
9. Wilson B, Cong C. Beyond the supply side: Use and impact of municipal open data in the US. Telemat Inf. 2021;101526:58. https://doi.org/10.1016/j.tele.2020.101526.
10. Demchenko Y, de Laat C, Membrey P. Defining architecture components of the Big Data Ecosystem. In: 2014 international conference on collaboration technologies and systems (CTS). IEEE, Minneapolis, Minnesota, USA 2014, pp. 104–112. https://doi.org/10.1109/CTS.2014.6867550. http://ieeexplore.ieee.org/document/6867550/.
11. Dawes SS, Vidiasova L, Parkhimovich O. Planning and designing open government data programs: an ecosystem approach. Govern Inf Q. 2016;33(1):15–27. https://doi.org/10.1016/j.giq.2016.01.003.
12. Magalhaes G, Roseira C, Manley L. Business models for open government data. In: ACM international conference proceeding series 2014-Janua, 2014, p 365–70. https://doi.org/10.1145/2691195.2691273.
13. Sutherland MK, Cook ME. Data-driven smart cities: a closer look at organizational, technical & data complexities. In: ACM international conference proceeding series Part. 2017, p 471–6. https://doi.org/10.1145/3085228.3085239.
14. Group DSR. Overview of the DDI Version 3. 0 conceptual model. Structural Reform Group. 2004. http://opendatafoundation.org/ddi/srg/Papers/DDIModel_v_4.pdf.
15. Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol Evol. 2012;27(2):85–93. https://doi.org/10.1016/j.tree.2011.11.016.
16. Arass ME, Tikito I, Souissi N. An audit framework for data lifecycles in a big data context. In: 2018 international conference on selected topics in mobile and wireless networking, MoWNeT. 2018. https://doi.org/10.1109/MoWNet.2018.8428883.
17. Arass ME, Tikito I, Souissi N. Data lifecycles analysis: Towards intelligent cycle. In: 2017 Intelligent Systems and Computer Vision (ISCV), pp. 1–8. IEEE, Fez, Morocco 2017. https://doi.org/10.1109/ISACV.2017.8054938. http://ieeexplore.ieee.org/document/8054938/.
18. Kitchenham: guidelines for performing systematic literature reviews in software engineering. Keele University, UK and University of Durham, UK 2007.
19. Cooper H, Hedges LV. Research synthesis as a scientific process. In: The handbook of research synthesis and meta-analysis. 2009, pp. 1–50.
20. Kitchenham B, Pretorius R, Budgen D, Pearl Brereton O, Turner M, Niazi M, Linkman S. Systematic literature reviews in software engineering—a tertiary study. Inf Softw Technol. 2010;52(8):792–805. https://doi.org/10.1016/j.infsof.2010.03.006.
21. Höchtl J, Parycek P, Schöllhammer R. Big data in the policy cycle: policy decision making in the digital era. J Organiz Comput Electr Commer. 2016;26(1–2):147–69. https://doi.org/10.1080/10919392.2015.1125187.
22. Nobubele AS, Mtsweni J. Big data privacy and security: a systematic analysis of current and future challenges. Pretoria: University of South Africa; 2016.
23. Khaloufi H, Abouelmehdi K, Beni-hssane A, Saadi M. Security model for Big Healthcare Data Lifecycle. Proc Comput Sci. 2018;141:294–301. https://doi.org/10.1016/j.procs.2018.10.199.
24. Immonen A, Kalaoja J. Requirements of an energy data ecosystem. IEEE Access. 2019;7:111692–708. https://doi.org/10.1109/ACCESS.2019.2933919.
25. Lukoianova T, Rubin VL. Veracity roadmap: is big data objective, truthful and credible? Adv Classif Res Online. 2014;24(1):4. https://doi.org/10.7152/acro.v24i1.14671.
26. Faroukhi AZ, El Alaoui I, Gahi Y, Amine A. Big data monetization throughout Big Data Value Chain: a comprehensive review. J Big Data. 2020;7(1):3. https://doi.org/10.1186/s40537-019-0281-5.

27. Pospiech M, Felden C. A Descriptive Big Data Model Using Grounded Theory. In: 2013 IEEE 16th international conference on computational science and engineering. IEEE, Sydney, NSW, Australia 2013, pp. 878–85. https://doi.org/10.1109/CSE.2013.132. http://ieeexplore.ieee.org/document/6755312/.
28. Ding L, Peristeras V, Hausenblas M. Linked open government data [Guest editors' introduction]. IEEE Intellig Syst. 2012;27(3):11–5. https://doi.org/10.1109/MIS.2012.56.
29. Lee D. Building an open data ecosystem. In: Proceedings of the 8th international conference on theory and practice of electronic governance. ACM, New York, NY, USA. 2014, pp. 351–60. https://doi.org/10.1145/2691195.2691258. https://dl.acm.org/doi/10.1145/2691195.2691258.
30. Misra D, Mishra A, Babbar S, Gupta V. Open Government Data Policy and Indian Ecosystems. In: Proceedings of the 10th international conference on theory and practice of electronic governance. ACM, New York, NY, USA. 2017, pp. 218–27. https://doi.org/10.1145/3047273.3047363. https://dl.acm.org/doi/10.1145/3047273.3047363.
31. organization S. Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions). In: Statista survey organization. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.
32. Boyi X, Da LX, Cai H, Xie C, Jingyuan H, Fenglin B. Ubiquitous data accessing method in IoT-based information system for emergency medical services. IEEE Trans Ind Inf. 2014;10(2):1578–86. https://doi.org/10.1109/TII.2014.2306382.
33. Bhat AZ, Ahmed I. Big data for institutional planning, decision support and academic excellence. In: 2016 3rd MEC international conference on big data and smart city (ICBDSC). IEEE, Muscat, Oman. 2016, pp. 1–5. https://doi.org/10.1109/ICBDSC.2016.7460353. http://ieeexplore.ieee.org/document/7460353/.
34. Barker TT. Finding pluto: an analytics-based approach to safety data ecosystems. Safety Health Work. 2021;12(1):1–9. https://doi.org/10.1016/j.shaw.2020.09.010.
35. Kaiser C, Stocker A, Viscusi G, Fellmann M, Richter A. Conceptualising value creation in data-driven services: the case of vehicle data. Int J Inf Manag. 2021;59: 102335. https://doi.org/10.1016/j.ijinfomgt.2021.102335.
36. Bernasconi A. Data quality-aware genomic data integration. Comput Methods Progr Biomed. 2021. https://doi.org/10.1016/j.cmpbup.2021.100009.
37. Adner R, Kapoor R. Value creation in innovation ecosystems: how the structure of technological interdependence affects firm performance in new technology generations. Strateg Manag J. 2010;31(3):306–33. https://doi.org/10.1002/smj.821.
38. Harrison TM, Pardo TA, Cook M. Creating open government ecosystems: a research and development agenda. Fut Intern. 2012;4(4):900–28. https://doi.org/10.3390/fi4040900.
39. Buteau S, Rao P, Mehta AK, Kadirvell V. Developing a framework to assess socio-economic value of open data in India. In: Proceedings of the 14th international symposium on open collaboration. ACM, New York, NY, USA. 2018, pp. 1–6. https://doi.org/10.1145/3233391.3233532. https://dl.acm.org/doi/10.1145/3233391.3233532.
40. Demchenko Y, Turkmen F, de Laat C, Blanchet C, Loomis C. Cloud based big data infrastructure: Architectural components and automated provisioning. In: 2016 international conference on high performance computing & simulation (HPCS). IEEE, Innsbruck, Austria. 2016, pp. 628–36. https://doi.org/10.1109/HPCSim.2016.7568394. http://ieeexplore.ieee.org/document/7568394/.
41. Grunzke R, Aguilera A, Nagel WE, Kruger J, Herres-Pawlis S, Hoffmann A, Gesing S. Managing Complexity in Distributed Data Life Cycles Enhancing Scientific Discovery. In: 2015 IEEE 11th international conference on e-Science. IEEE, USA. 2015, pp. 371–80. https://doi.org/10.1109/eScience.2015.72. http://ieeexplore.ieee.org/document/7304320/.
42. Sinaeepourfard A, Garcia J, Masip-Bruin X, Marin-Tordera E. A comprehensive scenario agnostic Data LifeCycle model for an efficient data complexity management. In: 2016 IEEE 12th international conference on e-Science (e-Science). IEEE, Baltimore, MD, USA. 2016, pp. 276–281. https://doi.org/10.1109/eScience.2016.7870909. http://ieeexplore.ieee.org/document/7870909/.
43. Sinaeepourfard A, Garcia J, Masip-Bruin X, Marín-Torder E. Towards a comprehensive data lifecycle model for big data environments. In: Proceedings of the 3rd IEEE/ACM international conference on big data computing, applications and technologies. ACM, New York, NY, USA. 2016, pp. 100–6. https://doi.org/10.1145/3006299.3006311. https://dl.acm.org/doi/10.1145/3006299.3006311.
44. Sinaeepourfard A, Garcia J, Masip-Bruin X, Marin-Tordera E, Yin X, Wang C. A data lifeCycle model for smart cities. In: 2016 international conference on information and communication technology convergence (ICTC), vol. 2. IEEE, Jeju, South Korea. 2016, pp. 400–5. https://doi.org/10.1109/ICTC.2016.7763506. http://ieeexplore.ieee.org/document/7763506/.
45. Zenggui O. Website data storage management during data lifecycle taking into account of time effect. In: 2008 IEEE 8th international conference on computer and information technology workshops. IEEE, Sydney, NSW, Australia. 2008, pp. 3–7. https://doi.org/10.1109/CIT.2008.Workshops.47. http://ieeexplore.ieee.org/document/4568470/.
46. Blazquez D, Domenech J. Big Data sources and methods for social and economic analyses. Technol Forecast Soc Change. 2018;13:99–113. https://doi.org/10.1016/j.techfore.2017.07.027.
47. Shameli-Sendi A. An efficient security data-driven approach for implementing risk assessment. J Inf Sec Appl. 2020. https://doi.org/10.1016/j.jisa.2020.102593.
48. Alshammari M, Simpson A. Personal data management: an abstract personal data lifecycle model. 2018, pp. 685–97. https://doi.org/fztc.
49. Möller K. Lifecycle models of data-centric systems and domains. Semantic Web. 2013;4(1):67–88. https://doi.org/10.3233/SW-2012-0060.
50. Rang W, Yang D, Cheng D, Wang Y. Data life aware model updating strategy for stream-based online deep learning. Trans Parall Distrib Syst. 2021;9219:1–12. https://doi.org/10.1109/tpds.2021.3071939.
51. IBM. Wrangling big data: fundamentals of data lifecycle management. IBM Managing data lifecycle. 2013.

52. Zuiderwijk A, Janssen M. Barriers and development directions for the publication and usage of open data: a socio-technical view. In: Open government: opportunities and challenges for public governance. Chap. Barriers. New York: Springer. 2014, pp. 115–35. https://doi.org/fztd.
53. Research data management team: data life cycle and data management planning. University of Essex, UK; 2013.
54. Sinaeepourfard A, Petersen SA. Distributed-to-centralized data management through data lifecycle models for zero emission neighborhoods. In: Communications in computer and information science, vol. 891. Cham: Springer. 2019, pp. 132–142. https://doi.org/fztf.
55. Faundeen JL, Burley TE, Carlino JA, Govoni DL, Henkel HS, Holl SL, Hutchison VB, Martin E, Montgomery ET, Ladino CC, Tessler S, Zolly LS. The United States geological survey science data lifecycle model. USA Govt. 2013. https://doi.org/10.3133/ofr20131265. http://pubs.usgs.gov/of/2013/1265/.
56. Ku M, Gil-Garcia JR. Ready for data analytics? In: Proceedings of the 19th annual international conference on digital government research: governance in the data age. New York: ACM. 2018, pp. 1–10. https://doi.org/10.1145/3209281.3209381. https://dl.acm.org/doi/10.1145/3209281.3209381.
57. Hardman L. Canonical processes of media production. In: Proceedings of the ACM workshop on multimedia for human communication from capture to convey—MHC '05. New York: ACM Press. 2005, p. 1. https://doi.org/10.1145/1099376.1099378. http://portal.acm.org/citation.cfm?doid=1837274.1837462http://portal.acm.org/citation.cfm?doid=1099376.1099378.
58. TD, PO. Public service Data Strategy 2019–2023. Government of Ireland. 2018. https://www.osi.ie/wp-content/uploads/2018/12/Public-Service-Data-Strategy-2019-2023.pdf.
59. Catteau, O., Vidal, P., Broisin, J.: A Generic Representation Allowing for Expression of Learning Object and Metadata Lifecycle. In: Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), vol. 2006, pp. 30–32. IEEE, Kerkrade, Netherlands (2006). https://doi.org/10.1109/ICALT.2006.1652357. http://ieeexplore.ieee.org/document/1652357/
60. El Arass M, Souissi N. Data Lifecycle: From Big Data to SmartData. In: 2018 IEEE 5th international congress on information science and technology (CiSt), vol. 2018. IEEE, Marrakech, Morocco. 2018, pp. 80–87. https://doi.org/10.1109/CIST.2018.8596547. https://ieeexplore.ieee.org/document/8596547/.
61. Khan N, Yaqoob I, Hashem IAT, Inayat Z, Mahmoud Ali WK, Alam M, Shiraz M, Gani A. Big data: survey, technologies, opportunities, and challenges. Sci World J. 2014;2014:1–18. doi: https://doi.org/10.1155/2014/712826.
62. Orenga-Roglá S, Chalmeta R. Framework for implementing a big data ecosystem in organizations. Commun ACM. 2018;62(1):58–65. https://doi.org/10.1145/3210752.
63. Crowston K, Qin J. A capability maturity model for scientific data management: evidence from the literature. Proc Am Soc Inf Sci Technol. 2011;48(1):1–9. https://doi.org/10.1002/meet.2011.14504801036.
64. Cuffe PK. Healy: data visualization: a practical introduction. IEEE Trans Profess Commun. 2019;6(3):310–1. https://doi.org/10.1109/TPC.2019.2922787.
65. Siddiqa A, Karim A, Gani A. Big data storage technologies: a survey. Front Inf Technol Elect Eng. 2017;18(8):1040–70. https://doi.org/10.1631/FITEE.1500441.
66. European Commission. Data strategy for digital transformation. European Commission.
67. Moiso C, Minerva R. Towards a user-centric personal data ecosystem The role of the bank of individuals' data. In: 2012 16th international conference on intelligence in next generation networks. IEEE, Berlin, Germany. 2012, pp. 202–9. https://doi.org/10.1109/ICIN.2012.6376027. http://ieeexplore.ieee.org/document/6376027/.
68. Loshin D. Using a machine learning data catalog to reboot data governance. Knowledge integrity, Inc. 2020. https://www.alation.com/wp-content/uploads/Reboot-Data-Governance-Whitepaper.pdf.
69. Moreno J, Fernandez EB, Serrano MA, Fernandez-Medina E. Secure development of big data ecosystems. IEEE Access. 2019;7:96604–19. https://doi.org/10.1109/ACCESS.2019.2929330.
70. International, D. DAMA-DMBOK data management body of knowledge, 2nd Ed. USA: Dama International; 2017.
71. Sekarhati DKS, Nefiratika A, Hidayanto AN, Budi NFA. Solikin: online travel agency (OTA) data maturity assessment: case study PT Solusi Awan Indonesia -"Flylist". In: 2019 international conference on information management and technology (ICIMTech). IEEE, Baltimore, MD, USA. 2019, pp. 492–7. https://doi.org/10.1109/ICIMTech.2019.8843728. https://ieeexplore.ieee.org/document/8843728/.
72. Georgiadis G, Poels G. Enterprise architecture management as a solution for addressing general data protection regulation requirements in a big data context: a systematic mapping study, vol. 19. New York: Springer. 2021, p. 313–62. https://doi.org/10.1007/s10257-020-00500-5.
73. Asih SN, Nabila R, Ismed IH, Fitriani WR, Hidayanto AN, Yudhoatmojo SB. Evaluation of data operations management maturity level using CMMI in a state-owned enterprise. In: 2019 5th international conference on computing engineering and design (ICCED), vol. 10. IEEE, Singapore. 2019, pp. 1–6. https://doi.org/10.1109/ICCED46541.2019.9161117. https://ieeexplore.ieee.org/document/9161117/.
74. Stewart J, Harte V, Sambrook S. What is theory? J Eur Ind Train. 2011;35(3):221–9. https://doi.org/10.1108/03090591111120386.
75. Shin D-H. Demystifying big data: anatomy of big data developmental process. Telecommun Policy. 2016;40(9):837–54. https://doi.org/10.1016/j.telpol.2015.03.007.
76. Oliveira MI, Barros Lima GF, Farias Loscio B. Investigations into data ecosystems: a systematic mapping study. Knowl Inf Syst. 2019;61(2):589–630. https://doi.org/10.1007/s10115-018-1323-6.
77. Clegg CW. Sociotechnical principles for system design. Appl Ergon. 2000;31(5):463–77. https://doi.org/10.1016/S0003-6870(00)00009-0.
78. Chyi Lee C, Yang J. Knowledge value chain. J Manag Dev. 2000;19(9):783–94. https://doi.org/10.1108/02621710010378228.
79. Haak E, Ubacht J, Van den Homberg M, Cunningham S, Van den Walle B. A framework for strengthening data ecosystems to serve humanitarian purposes. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age. ACM, New York. 2018, pp. 1–9. https://doi.org/10.1145/3209281.3209326. https://dl.acm.org/doi/10.1145/3209281.3209326.

80. Van Den Homberg M, Visser J, Van Der Veen M. Unpacking data preparedness from a humanitarian decision making perspective: Toward an assessment framework at subnational level. In: Proceedings of the international ISCRAM conference 2017. 2017, p 2–13.
81. Shah SIH, Vassilos Peristeras IM. Government big data ecosystem: definitions, types of data, actors and roles and the impact in public administrations. Journal of Data Information and Quality, 2021;13(2):1–25. https://doi.org/10.1145/3425709
82. Attard J, Orlandi F, Auer S. Data Value Networks: Enabling a New Data Ecosystem. In: Proceedings - 2016 IEEE/WIC/ACM international conference on web intelligence, WI 2016, Omaha. 2017, pp. 453–6. https://doi.org/10.1109/WI.2016.0073.
83. Zubcoff JJ, Vaquer L, Mazon JN, Macia F, Garrigos I, Fuster A, Carcel JV. The university as an open data ecosystem. Int J Design Nat Ecodyn. 2016;11(3):250–7. https://doi.org/10.2495/DNE-V11-N3-250-257.
84. Alex Ball. Review of data management lifecycle models. University of Bath. 2012. http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf.
85. Burton A, Treloar A. Designing for discovery and re-use?: The—ANDS Data Sharing Verbs' Approach to Service Decomposition. Int J Digit Curat. 2009;4(3):44–56.
86. Michota A, Katsikas S. Designing a seamless privacy policy for social networks. In: Proceedings of the 19th panhellenic conference on informatics, vol. 1. ACM, New York. 2015, pp. 139–43. https://doi.org/10.1145/2801948.2801998. https://dl.acm.org/doi/10.1145/2801948.2801998.
87. Lee SM, Hong S. An enterprise- wide knowledge management system infrastructure. Ind Manag Data Syst. 2002;102(1):17–25. https://doi.org/10.1108/02635570210414622.
88. Lin L, Liu T, Hu J, Zhang J. A privacy-aware cloud service selection method toward data life-cycle. In: 2014 20th IEEE international conference on parallel and distributed systems (ICPADS), vol. 2015. IEEE, Hsinchu, Taiwan. 2014, pp. 752–759. https://doi.org/10.1109/PADSW.2014.7097878. http://ieeexplore.ieee.org/document/7097878/.
89. Caíno-Lores S, Lapin A, Carretero J, Kropf P. Applying big data paradigms to a large scale scientific workflow: lessons learned and future directions. Fut Gener Comput Syst. 2020;110:440–52. https://doi.org/10.1016/j.future.2018.04.014.
90. Curation DCC, Model L. DCC Curation lifecycle model key elements of the DCC curation lifecycle model. 2015.
91. NIST Big Data Public Working Group. Definitions and taxonomies subgroup: NIST big data interoperability fframework: Volume 2, big data taxonomies. Technical report, National Institute of Standards and Technology, Gaithersburg. 2015. https://doi.org/10.6028/NIST.SP.1500-2. http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-2.pdfhttps://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-2.pdf.
92. Santos HDA, Oliveira MIS, Lima AB, Silva KM, Muniz RIVC, Lóscio BF. Investigations into data published and consumed on the Web: a systematic mapping study. J Braz Comput Soc. 2018;24(1):14. https://doi.org/10.1186/s13173-018-0077-z.
93. Lóscio BF, Oliveira MIS. Web publishing and consumption: concepts and challenges. SBBD. 2015.
94. Christopherson L, Mandal A, Scott E, Baldin I. Toward a data lifecycle model for NSF large facilities. In: Practice and experience in advanced research computing. ACM, New York. 2020, pp. 168–175. https://doi.org/10.1145/3311790.3396636. https://dl.acm.org/doi/10.1145/3311790.3396636.
95. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of Big Data Privacy. IEEE Access. 2016;4:1821–34. https://doi.org/10.1109/ACCESS.2016.2558446.
96. Viji D, Saravanan K, Hemavathi D. A journey on privacy protection strategies in big data. In: 2017 international conference on intelligent computing and control systems (ICICCS), vol. 2018. IEEE, Madurai, India. 2017, pp. 1344–7. https://doi.org/10.1109/ICCONS.2017.8250688. http://ieeexplore.ieee.org/document/8250688/.
97. Soltani Panah A, Yavari A, van Schyndel R, Georgakopoulos D, Yi X. Context-driven granular disclosure control for internet of things applications. IEEE Trans Big Data. 2019;5(3):408–22. https://doi.org/10.1109/TBDATA.2017.2737463.
98. Elmekki H, Chiadmi D, Lamharhar H. Open Government Data. In: Proceedings of the ArabWIC 6th annual international conference research track on—ArabWIC 2019. ACM Press, New York, New York, USA. 2019, pp. 1–6. https://doi.org/10.1145/3333165.3333180. http://dl.acm.org/citation.cfm?doid=3333165.3333180.
99. Arlene Fink. Conducting research literature reviews from the Internet to Paper. 2010, pp. 1–253.
100. Sandberg J, Alvesson M. Ways of constructing research questions: gap-spotting or problematization? Organization. 2011;18(1):23–44. https://doi.org/10.1177/1350508410372151.
101. Bouchrika I. How to write a research question: types, steps, and examples. 2021. https://www.guide2research.com/research/how-to-write-a-research-question.
102. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):6–8. https://doi.org/10.7326/ACPJC-1995-123-3-A12.
103. Doody O, Bailey ME. Setting a research question, aim and objective. Nurse Res. 2016;23(4):19–23. https://doi.org/10.7748/nr.23.4.19.s5.
104. Hulley Stephen B. Designing clinical research. Lippincott Williams, USA. 2007, pp. 1–367.
105. Lipowski EE. Developing great research questions. Am J Health-Syst Pharm. 2008;65(17):1667–70. https://doi.org/10.2146/ajhp070276.
106. Poojary S, Bagadia J. Reviewing literature for research: doing it the right way. Ind J Sex Transmitt Dis AIDS. 2014;35(2):85. https://doi.org/10.4103/0253-7184.142387.
107. Raszewski R, Goben AH, Bergren MD, Jones K, Ryan C, Steffen AD, Vonderheid SC. A survey of current practices in data management education in nursing doctoral programs. J Profess Nurs. 2021;37(1):155–62. https://doi.org/10.1016/j.profnurs.2020.06.003.
108. Heimstädt M, Saunderson F, Heath T. Conceptualizing Open Data Ecosystems: A timeline analysis of Open Data development in the UK. In: Proceedings of the international conference for E-democracy and open government (CeDEM2014).2014, p 1–11.

109. Jetten M, Simons E, Rijnders J. The role of CRIS's in the research life cycle. A case study on implementing a FAIR RDM policy at Radboud University, the Netherlands. Proc Comput Sci. 2019;146:156–65. https://doi.org/10.1016/j.procs.2019.01.090.
110. Attard J, Orlandi F, Scerri S, Auer S. A systematic review of open government data initiatives. Govern Inf Q. 2015;32(4):399–418. https://doi.org/10.1016/j.giq.2015.07.006.
111. Rahul K, Banyal RK. Data life cycle management in big data analytics. Proc Comput Sci. 2019;2020(173):364–71. https://doi.org/10.1016/j.procs.2020.06.042.
112. McKeever S. Understanding web content management systems: evolution, lifecycle and market. Ind Manag Data Syst. 2003;103(8–9):686–92. https://doi.org/10.1108/02635570310506106.
113. Freund GP, Fagundes PB, de Macedo DDJ. An analysis of blockchain and GDPR under the data lifecycle perspective. Mobile Netw Appl. 2021;26(1):266–76. https://doi.org/10.1007/s11036-020-01646-9.
114. Wang F, Harney J, Shipman G, Williams D, Cinquini L. Building a large scale climate data system in support of HPC environment. In: 2011 7th international conference on next generation web services practices. IEEE, Salamanca, Spain. 2011, pp. 380–5. https://doi.org/10.1109/NWeSP.2011.6088209. http://ieeexplore.ieee.org/document/6088209/.
115. Olsson U. Data management in telco networks: from costly duckling to profitable swan. In: 2010 14th International conference on intelligence in next generation networks. IEEE, Berlin, Germany. 2010, pp. 1–4. https://doi.org/10.1109/ICIN.2010.5640907. http://ieeexplore.ieee.org/document/5640907/.
116. Ku TY, Park WK, Choi H. Energy big data life cycle mechanism for renewable energy system (20172410100040), 2019–2020. 2019.
117. Jiang Y, Xu Y, Xu Q, Fang L, Lin C. Tobacco industry data security protection system. In: 2019 IEEE 4th international conference on computer and communication systems, ICCCS. 2019, p 159–163. https://doi.org/10.1109/CCOMS.2019.8821674.
118. Chen D, Zhao H. Data security and privacy protection issues in cloud computing. In: Proceedings—2012 international conference on computer science and electronics engineering, ICCSEE. 2012, p 647–51. https://doi.org/10.1109/ICCSEE.2012.193.
119. Dqg K, Azer MA. Cloud computing privacy issues, challenges and solutions. https://doi.org/10.1109/ICCES.2017.8275295.
120. Cao H, Wachowicz M, Renso C, Carlini E. Analytics everywhere: generating insights from the internet of things. IEEE Access. 2019;7:71749–69. https://doi.org/10.1109/ACCESS.2019.2919514.
121. Peng J, Huang X, Li M, Zhang J, Zhang Y, Gao N. Differential Attribute Desensitization System for Personal Information Protection. In: 2019 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, Leicester, UK. 2019, pp. 1243–8. https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00231. https://ieeexplore.ieee.org/document/9060310/.
122. Kumar S, Singh M. Big data analytics for healthcare industry: impact, applications, and tools. Big Data Min Anal. 2019;2(1):48–57. https://doi.org/10.26599/BDMA.2018.9020031.
123. Solanas A, Casino F, Batista E, Rallo R. Trends and challenges in smart healthcare research: A journey from data to wisdom. In: RTSI 2017—IEEE 3rd international forum on research and technologies for society and industry, conference proceedings. 2017, p 17–22. https://doi.org/10.1109/RTSI.2017.8065986.
124. Hasan R, Myagmar S, Lee AJ, Yurcik W. Toward a threat model for storage systems. In: StorageSS'05—Proceedings of the 2005 ACM workshop on storage security and survivability, 2005, p 94–102. https://doi.org/10.1145/1103780.1103795.
125. Khan HR, Chang H-C, Kim J. Unfolding research data services. 2018, p 353–4. https://doi.org/10.1145/3197026.3203887.
126. Borgman CL, Wallis JC, Mayernik MS, Pepe A. Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. In: Proceedings of the ACM international conference on digital libraries. 2007, p 269–77. https://doi.org/10.1145/1255175.1255228.
127. Santos MY, Oliveira SA, Andrade C, ValeLima F, Costa E, Costa C, Martinho B, Galvão J. A Big Data system supporting Bosch Braga Industry 4.0 strategy. Int J Inf Manag. 2017;37(6):750–60. https://doi.org/10.1016/j.ijinfomgt.2017.07.012.
128. Zambetti M, Pinto R, Pezzotta G. Data lifecycle and technology-based opportunities in new product service system offering towards a multidimensional framework. Proc CIRP. 2019;83:163–9. https://doi.org/10.1016/j.procir.2019.02.135.
129. Khan S, Liu X, Shakil KA, Alam M. A survey on scholarly data: from big data perspective. Inf Process Manag. 2017;53(4):923–44. https://doi.org/10.1016/j.ipm.2017.03.006.
130. Tao F, Qi Q, Liu A, Kusiak A. Data-driven smart manufacturing. J Manufact Syst. 2018;48:157–69. https://doi.org/10.1016/j.jmsy.2018.01.006.
131. Lyon L, Jeng W, Mattern E. Developing the tasks-toward-transparency (T3) model for research transparency in open science using the lifecycle as a grounding framework. Libr Inf Sci Res. 2020. https://doi.org/10.1016/j.lisr.2019.100999.
132. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data. 2018;5(1):1. https://doi.org/10.1186/s40537-017-0110-7.
133. Levitin AV, Redman TC. A model of the data (life) cycles with application to quality. Inf Softw Technol. 1993;35(4):217–23. https://doi.org/10.1016/0950-5849(93)90069-F.
134. Michener WK, Allard S, Budden A, Cook RB, Douglass K, Frame M, Kelling S, Koskela R, Tenopir C, Vieglais DA. Participatory design of DataONE-Enabling cyberinfrastructure for the biological and environmental sciences. Ecol Inf. 2012;11:5–15. https://doi.org/10.1016/j.ecoinf.2011.08.007.
135. Muthy S, Sookram T, Gobin-Rahimbux B. Big data analytics life cycle for social networks' posts. In: Advances in intelligent systems and computing, vol. 863. Springer, Singapore. 2019, pp. 379–88. https://doi.org/fzt5.

136. OECD. Data-driven innovation for growth and well-being. OECD. 2014. https://www.oecd.org/sti/inno/data-driven-innovation-interim-synthesis.pdf.
137. Kilov H. From semantic to object-oriented data modeling. In: Systems integration '90. Proceedings of the first international conference on systems integration. IEEE Comput. Soc. Press, Morristown, NJ, USA. 1990, pp. 385–93. https://doi.org/10.1109/ICSI.1990.138704. http://ieeexplore.ieee.org/document/138704/.
138. Staab S, Studer R, Schnurr H-P, Sure Y. Knowledge processes and ontologies. IEEE Intellig Syst. 2001;16(1):26–34. https://doi.org/10.1109/5254.912382.
139. Qingqing T, Mengting N, Juan W. An Intelligent Recommendation Mobile Application Privacy Risk Evaluation Method Based on Optimized SVM. In: Proceedings of 2020 IEEE 4th information technology, networking, electronic and automation control conference, ITNEC. 2020, p 2486–91. https://doi.org/10.1109/ITNEC48623.2020.9085180.
140. Cheng X, Hu C, Li Y, Lin W, Zuo H. Data Evolution Analysis of Virtual DataSpace for Managing the Big Data Lifecycle. In: 2013 IEEE international symposium on parallel & distributed processing, workshops and Phd Forum. IEEE, Cambridge, MA, USA. 2013, pp. 2054–63. https://doi.org/10.1109/IPDPSW.2013.57. http://ieeexplore.ieee.org/document/6651110/.
141. Cao J, Diao X, Jiang G, Du Y. Data lifecycle process model and quality improving framework for TDQM Practices. In: 2010 international conference on E-product E-service and E-entertainment. IEEE, Henan, China. 2010, pp. 1–6. https://doi.org/10.1109/ICEEE.2010.5661270. http://ieeexplore.ieee.org/document/5661270/.
142. Yu, X., Wen, Q.: A view about cloud data security from data life cycle. In: 2010 international conference on computational intelligence and software engineering, CiSE 2010 (4072020). 2010, p 3. https://doi.org/10.1109/CISE.2010.5676895
143. Jennifer L. Bauer LP, Haley St.. Dennis: data brokers and human rights (Big Data, Big Business). Institute for Human Rights and Business (IHRB). 2016.
144. Tianfield H. Cyber security situational awareness. In: 2016 IEEE international conference on internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData). IEEE, Chengdu, China. 2016, pp. 782–7. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.165. http://ieeexplore.ieee.org/document/7917193/.
145. Millard DE, Tao F, Doody K, Woukeu A, Davis HC. The knowledge life cycle for e-learning. Int J Continu Eng Educ Life-Long Learn. 2006;16(1/2):110. https://doi.org/10.1504/IJCEELL.2006.008921.
146. Demchenko Y, Ngo C, de Laat C, Membrey P, Gordijenko D. Big security for big data: addressing security challenges for the big data infrastructure, vol. 8425 LNCS. 2014, p 76–94. https://doi.org/fzt6.
147. Corujo L, da Silva CG, Revez J. Digital curation and costs: approaches and perceptions (Dcc). 2016, p 277–84. https://doi.org/10.1145/3012430.3012529.
148. Ahn S, Oh H, Kim HJ, Choi JK. Data lifecycle and tagging for internet of things applications. In: Lecture notes in computer science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9992 LNAI. 2016, pp. 691–5. https://doi.org/fzt7. http://link.springer.com/10.1007/978-3-319-50127-7_61.
149. Kiourtis A, Mavrogiorgou A, Kyriazis D, Maglogiannis I, Themistocleous M. Towards data interoperability: turning domain specific knowledge to agnostic across the data lifecycle. In: 2016 30th international conference on advanced information networking and applications workshops (WAINA). IEEE, Crans-Montana, Switzerland. 2016, pp. 109–114. https://doi.org/10.1109/WAINA.2016.69. http://ieeexplore.ieee.org/document/7471182/.
150. Ma X, Fox P, Rozell E, West P, Zednik S. Ontology dynamics in a data life cycle: challenges and recommendations from a Geoscience Perspective. J Earth Sci. 2014;25(2):407–12. https://doi.org/10.1007/s12583-014-0408-8.
151. ITU: Regional Classifications by ITU 2021. https://www.itu.int/en/ITU-D/Statistics/Pages/definitions/regions.aspx.
152. Ministry of Education. Action plan—implementing DataStrategy@EC. European Commission 2018.
153. Demchenko Y, Grosso P, de Laat C, Membrey P. Addressing big data issues in Scientific Data Infrastructure. In: 2013 International conference on collaboration technologies and systems (CTS). IEEE, San Diego, CA, USA. 2013, pp. 48–55. https://doi.org/10.1109/CTS.2013.6567203. http://ieeexplore.ieee.org/document/6567203/.
154. Xianglan L. Digital construction of coal mine big data for different platforms based on life cycle. In: 2017 IEEE 2nd international conference on big data analysis, ICBDA. 2017, p 456–9. https://doi.org/10.1109/ICBDA.2017.8078862.
155. Kaufmann M. Towards a reference model for big data management. 2016.
156. Gartner. What is a big data. Gartner Publications. 2019. https://www.gartner.com/en/information-technology/glossary/big-data%0A.
157. Heimstädt M. "The Institutionalization of Digital Openness". In: Proceedings of The international symposium on open collaboration. ACM, New York, NY, USA. 2014, pp. 1–2. https://doi.org/10.1145/2641580.2641626. http://dl.acm.org/citation.cfm?doid=2641580.2641626 https://dl.acm.org/doi/10.1145/2641580.2641626.
158. Gajbe SB, Tiwari A, Gopalji Singh RK. Evaluation and analysis of data management plan tools: a parametric approach. Inf Process Manag. 2021. https://doi.org/10.1016/j.ipm.2020.102480.
159. Tikito I, Souissi N. Data Collect Requirements Model. In: Proceedings of the 2nd international conference on big data, cloud and applications, Part F1294. ACM, New York, NY, USA. 2017, pp. 1–7. https://doi.org/10.1145/3090354.3090358. http://dl.acm.org/doi/10.1145/3090354.3090358.
160. Divakar M, Shrikant Khupat SJ. Introduction to big data architecture,.2017, p 1–14.
161. Nguyen DC, Cheng P, Ding M, Lopez-Perez D, Pathirana PN, Li J, Seneviratne A, Li Y, Poor HV. Enabling AI in future wireless networks: a data life cycle perspective. IEEE Commun Surv Tutor. 2021;23(1):553–95. https://doi.org/10.1109/COMST.2020.3024783.
162. Hai R, Geisler S, Quix C. Constance. In: Proceedings of the 2016 international conference on management of data, vol. 26. ACM, New York, NY, USA. 2016, pp. 2097–2100. https://doi.org/10.1145/2882903.2899389. https://dl.acm.org/doi/10.1145/2882903.2899389.
163. Ambrosio LM, Marques P, David JMN, Braga R, Ribeiro Dantas MA, Stroele V, Campos F. An approach to support data integration in a scientific software ecosystem platform. In: 2019 IEEE 23rd international conference on computer supported cooperative work in design (CSCWD). IEEE, Porto, Portugal. 2019, pp. 39–44. https://doi.org/10.1109/CSCWD.2019.8791499. https://ieeexplore.ieee.org/document/8791499/.

164. Smith HA, McKeen JD. Developments in practice XXX: master data management: salvation or snake oil? Commun Assoc Inf Syst. 2008;23:245. https://doi.org/10.17705/1CAIS.02304.
165. Biesialska K, Franch X, Muntés-Mulero V. Big Data analytics in Agile software development: a systematic mapping study. Inf Softw Technol. 2021. https://doi.org/10.1016/j.infsof.2020.106448.
166. Thomas L, Gougeaud S, Deniel P. Predicting file lifetimes for data placement in multi-tiered storage systems for HPC. In: Proceedings of the workshop on challenges and opportunities of efficient and performant storage systems (CHEOPS '21). Association for computing machinery, New York, NY, USA, New York, NY, USA. 2021, p 1–9. https://doi.org/10.1145/3439839.3458733.
167. Sawadogo P, Darmont J. On data lake architectures and metadata management. J Intellig Inf Syst. 2021;56(1):97–120. https://doi.org/10.1007/s10844-020-00608-7.
168. María J, Edward C. New horizons for a data-driven economy. Springer, Cham. 2016, pp. 1–312. https://doi.org/10.1007/978-3-319-21569-3. http://link.springer.com/10.1007/978-3-319-21569-3.
169. Zhang X, Wang Y. Research on intelligent medical big data system based on Hadoop and blockchain. Eur J Wirel Commun Netw. 2021;2021:124. https://doi.org/10.1186/s13638-020-01858-3.
170. Zuiderwijk A, Janssen M, Van De Kaa G, Poulis K. The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments. Information Polity. 2016;21(3):223–36. https://doi.org/10.3233/IP-160391.
171. Gade M, Koolen M, Hall M, Bogers T, Petras V. A Manifesto on resource re-use in interactive information retrieval. In: CHIIR 2021—Proceedings of the 2021 conference on human information interaction and retrieval, Canberra, ACT, Australia. 2021, pp. 141–9. https://doi.org/10.1145/3406522.3446056.
172. Immonen A, Palviainen M, Ovaska E. Requirements of an open data based business ecosystem. IEEE Access. 2014;2:88–103. https://doi.org/10.1109/ACCESS.2014.2302872.
173. Acharya S, Park HW. Open data in Nepal: a webometric network analysis. Qual Quant. 2017;51(3):1027–43. https://doi.org/10.1007/s11135-016-0379-1.
174. Abebe R, Aruleba K, Birhane A, Kingsley S, Obaido G, Remy SL, Sadagopan S. Narratives and counternarratives on data sharing in Africa. In: FAccT 2021—Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, Canada. 2021, p 329–41. https://doi.org/10.1145/3442188.3445897. arXiv:2103.01168
175. Chattapadhyay S: Access and use of government data by research and advocacy organisations in India. In: Proceedings of the 8th international conference on theory and practice of electronic governance. 2014. ACM, New York, NY, USA. 2014, pp. 361–4. https://doi.org/10.1145/2691195.2691262. https://dl.acm.org/doi/10.1145/2691195.2691262.
176. data Portal E. Open data Goldbook for data managers and data holders. In: European Commission. 2018, p 1–80.
177. Opengovdata: The 8 Principles of Open Government Data (OpenGovData.org). 2021. https://opengovdata.org/. Accessed 6 Mar 2021.
178. BOUTEILLER S. How to manage corporate data to create value—CIGREF. CIGREF.2014.
179. Mouzakitis S, Papaspyros D, Petychakis M, Koussouris S, Zafeiropoulos A, Fotopoulou E, Farid L, Orlandi F, Attard J, Psarras J. Challenges and opportunities in renovating public sector information by enabling linked data and analytics. Inf Syst Front. 2017;19(2):321–36. https://doi.org/10.1007/s10796-016-9687-1.
180. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, Huebner M, Schmidt B, Sauerbrei W, Richter A. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol. 2021;21(1):1–15. https://doi.org/10.1186/s12874-021-01252-7.
181. Haider A, Haider W. Improving engineering asset lifecycle data quality: setting the rules. 2013 Proceedings of PICMET 2013: technology management in the IT-driven services. 2013, p 1200–1206.
182. Meurisch C, Mühlhäuser M. Data protection in AI services. ACM Comput Surv. 2021;54(2):1–38. https://doi.org/10.1145/3440754.
183. Cumbley R, Church P. Is, Big Data creepy? Comput Law Secur Rev. 2013;29(5):601–9. https://doi.org/10.1016/j.clsr.2013.07.007.
184. Liang W, Ji N. Privacy challenges of IoT-based blockchain: a systematic review. Clust Comput. 2021;1:1–19. https://doi.org/10.1007/s10586-021-03260-0.
185. Jang K, Kim WJ. Development of data governance components using DEMATEL and content analysis. J Supercomput. 2020;14:87. https://doi.org/10.1007/s11227-020-03405-9.
186. Otto B. Data governance. Bus Inf Syst Eng. 2011;3(4):241–4. https://doi.org/10.1007/s12599-011-0162-8.
187. Suicimezov N, Georgescu MR. IT governance in cloud. Proc Econ Fina. 2014;15(14):830–5. https://doi.org/10.1016/S2212-5671(14)00531-0.

Publisher's Note