Knowledge and Information Systems
https://doi.org/10.1007/s10115-018-1323-6

SURVEY PAPER

Investigations into Data Ecosystems: a systematic mapping study

Marcelo Iury S. Oliveira[1,2] · Glória de Fátima Barros Lima[1] · Bernadette Farias Lóscio[1]

Received: 2 August 2017 / Revised: 20 August 2018 / Accepted: 20 October 2018

Abstract

Data Ecosystems are socio-technical complex networks in which actors interact and collaborate with each other to find, archive, publish, consume, or reuse data as well as to foster innovation, create value, and support new businesses. While the Data Ecosystem field is thus arguably gaining in importance, research on this subject is still in its early stages of development. Up until now, not many academic papers related to Data Ecosystems have been published. Furthermore, to the best of our knowledge, there has been no systematic review of the literature on Data Ecosystems. In this study, we provide an overview of the current literature on Data Ecosystems by conducting a systematic mapping study. This study is intended to function as a snapshot of the research in the field and by doing so identifies the different definitions of Data Ecosystem and analyzes the evolution of Data Ecosystem research. The studies selected have been classified into categories related to the study method, contribution, research topic, and ecosystem domains. Finally, we analyze how Data Ecosystems are structured and organized, and what benefits can be expected from Data Ecosystems and what their limitations are.

Keywords Data Ecosystem · Data consumption · Data provision · Systematic mapping

This paper was partially supported by funds from the following Brazilian funding agencies: the Pernambuco State Science and Technology Support Foundation (FACEPE) and the National Institute of Science and Technology for Software Engineering (INES). Marcelo Iury also thanks the National Committee of Technological and Scientific Development (CNPq) and the Coordination for the Improvement of Higher Level Personnel (CAPES), for granting him a PhD fellowship. The authors would also like to thank the colleagues of the Aladin research group for their input to this study.

✉ Marcelo Iury S. Oliveira
miso@cin.ufpe.br; marcelo.iury@ufrpe.br

Glória de Fátima Barros Lima
gfabl@cin.ufpe.br

Bernadette Farias Lóscio
bfl@cin.ufpe.br

[1] Center for Informatics, Federal University of Pernambuco, Recife, PE, Brazil

[2] Academic Unit of Serra Talhada, Federal Rural University of Pernambuco, Serra Talhada, PE, Brazil

Published online: 01 January 2019

1 Introduction

The highly rapid development of networks, the Internet of things (IoT), and Web-related technologies opens up possibilities for capturing, storing, and analyzing collections of data [8,11,46]. Governments, research institutions, and individuals are producing and making available large amounts of data on a variety of platforms (e.g., the Web, sensor-based applications, and social networks) [11,46]. An increasing number of individuals are recognizing the importance of data and as a consequence are setting up platforms for publishing, trading, or even selling data [7,55].

According to [63], in the majority of these cases, the current basic model for the provision and usage of data is a one-way street. There is no feedback loop between data users and data providers; i.e., data users do not share data and knowledge back to their data providers. In an ideal scenario, data users should share back their cleaned and integrated data, for example. Data users should also be able to give their contribution by flagging errors, or by submitting corrections. Indeed, all data users and data providers should be able to collaborate. In order to unlock the potential benefits of sharing data, a Data Ecosystem needs to be established [75]. Such ecosystem should be built on collaboration and coordination between various stakeholders, including public and private organizations, development partners, and end users.

In this context, a Data Ecosystem may be defined as a complex socio-technical network that enables collaboration between autonomous actors in order to explore data [41,63,75, 87]. Such ecosystems provide an environment for creating, managing, and sustaining data sharing initiatives [29,41,75,87], such as Smart Cities [2], open data [41], and Scientific Data Communities [43].

The emergence of Data Ecosystems has been driven by several factors, including the emergence of digital technologies and political/institutional initiatives. For instance, most Data Ecosystems have been mainly driven by the open data movement and Open Government Data (OGD) programs, which call for the free use, reuse, and redistribution of data by anyone [25]. Several governments have already launched Open Data Portals to stimulate and promote open data production and consumption [13]. Improvements in the technology (e.g., mobile Internet or technology) and trends in the technology (e.g., social media or mobile apps) also have been driving private and public organizations to publish data as well as to integrate their services with external data.

While Data Ecosystems are thus arguably gaining in importance, research into Data Ecosystems is still in its seminal stages. Up until now, not many academic papers related to the Data Ecosystem field have been published. In most cases they are focused on some component or technology that reflects only a small fragment of the whole research area. The same can be said of Data Ecosystems theory which does not yet provide a full conceptual basis for further studies into the research field. The terminology and definitions for Data Ecosystem vary greatly. This diversity raises a pressing problem for the development of a clear understanding about the new opportunities and emergent challenges in exploiting Data Ecosystems. Accurate definitions are required in order to reach a shared understanding of what Data Ecosystems embody.

A recent study [29] reviews the literature to analyze the agenda for a Government Data Ecosystem and also analyzes some principles that characterize how it should function. More recently, the authors of [87] review the literature in order to map the essential elements of Open Data Ecosystems. They found that an Open Data Ecosystem is characterized by multiple interdependent socio-technical levels and dimensions. The authors also discuss definitions, challenges, and barriers of Open Data Ecosystems. These two studies already provide an important understanding of how Data Ecosystems have been developed, although they focused on specific aspects of the open data concept or their analysis of the state of the art was somewhat narrow. Moreover, the coverage and the depth of analysis of the literature provided by these review articles are not as great as in our review.

Therefore, systematic and holistic reviews are necessary in order to provide further insights into the current state of Data Ecosystems as well as on the overall development of the topic, which can form the basis for shaping future research. To the best of our knowledge, until now there has been no systematic literature review on Data Ecosystems that seeks to map the state of the art, while also identifying research gaps and expected benefits.

In this study, we provide an overview of the current literature on Data Ecosystems by conducting a systematic mapping study. Systematic mapping is a protocol-driven methodology for reviewing and synthesizing a research area [38]. A systematic mapping study typically provides an overview of the research reported in the field and identifies possible issues related to the existing literature. This study is intended to function as a snapshot of the research in the Data Ecosystem area by (i) identifying and analyzing the different definitions of Data Ecosystem, (ii) analyzing the evolution of Data Ecosystem research, (iii) classifying existing studies into categories such as type of contribution and the problem studied, (iv) defining and analyzing how Data Ecosystems are structured and organized, (v) analyzing the benefits that can be expected from Data Ecosystems and what the limitations are, and (vi) presenting a landscape for a Data Ecosystem by discussing how a Data Ecosystem is characterized and how it works.

The remainder of the study is organized as follows: Sect. 2 gives a description of the research methodology used in our study. Section 3 analyzes the results. A discussion on the Data Ecosystem landscape is presented in Sect. 4. Section 5 suggests some directions for future research. Finally, some conclusions are drawn in Sect. 6.

2 Research approach

This study takes the form of a review of the literature on Data Ecosystems. It sets out to provide an overview of this field by undertaking a systematic mapping study, which is a protocol-driven review and synthesis of data focusing on a topic or on related key questions. The protocol specifies the methods used to guide the selection of the corpus of studies, the extraction of data, and the analysis and interpretation of the results. Moreover, a review protocol reduces the possibility of researcher bias and provides better transparency and reproducibility [38,60]. Unlike the conventional literature review, systematic mapping is more suitable for dealing with broad and poorly defined areas [38,60]. In addition, a systematic mapping is also more readily able to answer broader exploratory questions (e.g., *what do we know about topic "T"?*) [4].

Figure 1 illustrates the adopted protocol, whereas the individual steps are explained in Sects. 2.1–2.4. The applied review protocol is based on the guidelines of [38] and is focused on a set of research questions.

2.1 Research questions

The purpose of this systematic literature review is to provide an overview of the research reported in the field of Data Ecosystems. For this, the following general research question was defined: *RQ: What is the current state of the art regarding Data Ecosystems research?*

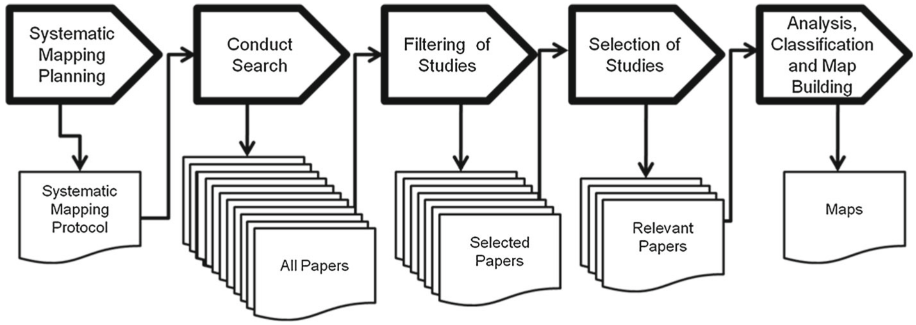


Fig. 1 Systematic literature mapping process. Adapted from: [44]

We also defined 13 specific research questions to guide and structure the extraction and analysis of research data, and the synthesis of evidence. In particular, these specific research questions are used to categorize and quantify the key contributions and the evolution of research concerning Data Ecosystems, as well as to discover issues that still need to be explored and the limitations of existing studies.

- RQ1: What has been the evolution in the number of Data Ecosystems studies over the years?
- RQ2: What level of focus (full and partial) do the studies dedicate to Data Ecosystem research?
- RQ3: Who are the main publishers of Data Ecosystem studies?
- RQ4: Which individuals and organizations are the most active in Data Ecosystem research?
- RQ5: What types of contributions are reported by the studies?
- RQ6: Which topics or themes have been addressed by Data Ecosystem studies?
- RQ7: Which research methods have been used in Data Ecosystem studies?
- RQ8: How is the term "Data Ecosystem" defined?
- RQ9: What theoretical foundations are adopted by Data Ecosystems studies?
- RQ10: What are the main roles of actors in a Data Ecosystem?
- RQ11: How are Data Ecosystems structured and organized?
- RQ12: What are the main domains in which Data Ecosystems have been and are being developed?
- RQ13: What is currently known about the benefits and limitations of Data Ecosystems?

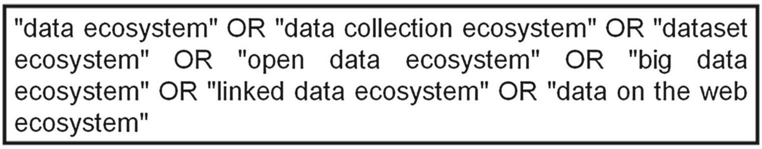
2.2 Data sources and search strategy

The strategy for collecting the relevant literature was to undertake a keyword search in five academic libraries (Table 1). The automated search process which took place in March 2017 retrieved 244 studies. The keyword query string (Fig. 2) consists of synonyms of the search term "Data Ecosystem" defined according to our expertise. Also, we used wildcard characters to capture the plural and singular forms of the keywords. The query string was intentionally kept simple, with no predicate, which enabled us to extract the maximum number of studies containing the terms.

From the initial set of 244 studies, we selected studies that mentioned concepts, theories, guidelines, discussions, lessons learned, and reports about the Data Ecosystem field (inclusion

Table 1 Sources and number of studies

Academic library and snowballing	Studies retrieved	Studies excluded	Studies included
IEEE	27	22	5
ACM	20	19	1
ScienceDirect	18	14	4
Scopus	105	93	12
Springer	74	73	1
Snowballing	6	0	6
Total	250	221	29



"data ecosystem" OR "data collection ecosystem" OR "dataset ecosystem" OR "open data ecosystem" OR "big data ecosystem" OR "linked data ecosystem" OR "data on the web ecosystem"

Fig. 2 Keyword-based query string used to automate for search studies. *Source*: Authors

criteria). We excluded studies that fell into any of the six exclusion criteria (EC), which were as follows:

- EC1: The study is not peer-reviewed (e.g., presentation slides, extended abstracts, invited papers, keynote speech, workshop reports, books);
- EC2: The study contains less than one page;
- EC3: The study is not written in English;
- EC4: The study is not accessible on the Web;
- EC5: The study does not present any type of findings or discussion about Data Ecosystems;
- EC6: The study is duplicated.

2.3 Selection of studies

The automatic search for and selection of studies was led by the first and second authors, who conducted the steps presented in Fig. 1. Many of the retrieved studies were eliminated in the filtering step, during which the researchers evaluated the studies by looking at the title, abstract, and venue information. After the exclusion criteria were applied, 48 potentially relevant studies remained.

After the filtering of studies step, the researchers worked on the selection step. Initially, the 48 potentially relevant studies were analyzed by both researchers, each of whom worked independently. The researchers applied the inclusion and exclusion criteria on the complete texts of potentially relevant studies. The differences were resolved in a meeting at which all authors reached a consensus. Duplicated studies were excluded in this step. However, for studies published in more than one academic library, all versions were reviewed for the purpose of data extraction.

Moreover, we also applied the snowballing technique to find the hidden population of studies, i.e., those studies that were not returned by our automatic search process but that could be interesting for the research. According to [34], snowballing, when compared to

Table 2 Rationale for excluding papers

Rationale	No. of studies
Do not address Data Ecosystems	154
Duplicated	40
Paper is not an academic work or peer-reviewed work	21
Not in English language	4
Document body no longer than one page	2
Total	221

automated searches in databases conducted in systematic literature searches, is less costly and brings less “noise” (articles that will not be selected). In particular, we used the backward snowballing approach that looks for new studies in the references of our list of studies. The whole selection process considered 29 studies as relevant ones for the data extraction and analysis (cf. Table 1).

During the filtering and selection phases, we kept track of the rationale for each exclusion, as shown in Table 2. Although all studies mention the term “Data Ecosystem,” a significant number of them do not investigate Data Ecosystems narrowly, nor do they take a broad view. In fact, a large number of the studies used the term “Data Ecosystem” only once throughout the text or used the term as a keyword so that the study would be flagged during a search. Another set of excluded studies present a solution for data publication or data consumption, but they do not, for instance, focus on the relationships between data providers and other actors within an ecosystem. Moreover, 40 studies are indexed by two or more academic libraries. In most of the cases of duplication, we chose the study from the Scopus library, since this library retrieved the largest number of studies.

2.4 Data extraction and synthesis

The data were extracted using an electronic spreadsheet in Google Spreadsheets™. Each study was analyzed, and the information collected was recorded on a form. Each researcher worked independently to extract data from the whole universe of studies. In the end, the most experienced researcher reviewed and, if necessary, revised all the data extracted by analyzing the studies again. This revision activity is meant to improve the accuracy of the extraction process and, therefore, the reliability of the results. Also, conflicts during the extraction phase were discussed and resolved by consensus at a meeting with the three authors of this study.

During the data extraction, classification schemes and categories for the selected studies were created. Our classification schemes started with an initial version based on previous studies, such as the taxonomy proposed by Shaw [65]. The classification schemes were reviewed during data extraction and data analysis, and this resulted in adding new categories or merging or splitting existing ones. The iterative refinement process was finalized when the schemes were able to consistently categorize all the relevant studies selected. We used mind-maps to assist the data analysis and to refine the classification schemes, which were included in the last step of our process (cf. Fig. 1).

The classification schema consists of four facets:

- Research focus (cf. Table 3): to distinguish between studies entirely devoted to Data Ecosystems and those that have a more narrow perspective;
- Contribution type (cf. Table 3): to map the different types of study outcomes;

Table 3 Classification scheme: research type facet and contribution type facet

Category	Description
(a) Research focus facet	
Full	A study entirely devoted to Data Ecosystems research
Partial	A study that has a narrow perspective and discusses some specific aspects of a Data Ecosystem
(b) Contribution type facet (adapted from [65])	
Method	A study that searches for a general solution to a problem
New tool	A study that proposes a new tool created or implemented by applying some method or technique, which may be more effective than existing tools or used in combination with other tools
Empirical model	A study that proposes an empirical predictive model based on observed data
Descriptive model	A study that proposes a structure or taxonomy for a problem area, such as architectural style, framework, or design pattern; non-formal domain analysis, well-grounded checklists, well-argued informal generalizations, guidance for integrating other results, well-organized interesting observations
Report	A study documenting knowledge and experience obtained, rules of thumb or checklists
Analysis	A study that analyzes a study object with regard to a structure or taxonomy, method, framework, or any set of evaluation criteria

- Research theme (cf. Table 4): to describe the topic or problem addressed in the study;
- Research method (cf. Table 4): to map the research method used in the study undertaken.

It is important to remark that not all the research questions require the use of a classification scheme. For instance, the answers to RQ8-RQ14 were constructed by analyzing the data extracted from each study and synthesizing them using thematic analysis and qualitative coding techniques.

The results from the data extraction phase were integrated on spreadsheets, which were also used to generate mind-maps and tables. All descriptive information was calculated and organized using Google Spreadsheets™. The data extracted from each study were synthesized using thematic analysis and qualitative coding techniques. The synthetic data were organized into mind-maps. Each row of a spreadsheet as well as the synthetic mind-maps were reviewed more than once by at least two researchers.

2.5 Threats to validity

The threats to validity in this systematic mapping were as follows:

Selection of relevant studies We defined research questions in advance and devised the inclusion and exclusion criteria in order to ensure an unbiased selection. However, some important fundamental works might be excluded. This threat was mitigated by selecting different terms to represent Data Ecosystems as well as by not using any filter predicate in the search query string used in the automatic searches.

Missing relevant studies The search for studies was conducted in five search engines, even though it is possible we missed some relevant studies. Nevertheless, this threat was mitigated by selecting the search engines which have been considered the most relevant scientific

Table 4 Classification scheme: research theme facet and research method facet

Category	Description
(c) Research theme facet	
Functions and features of Data Ecosystems	A study that describes key features and functions of Data Ecosystems
Describing and modeling Data Ecosystems	A study that describes Data Ecosystems in order to achieve a more thorough understanding of Data Ecosystems
Analysis and evaluation of Data Ecosystem	A study that analyzes and evaluates a Data Ecosystem initiative/experience
Structuring and shaping of Data Ecosystems	A study that describes how Data Ecosystems are organized and structured
Solutions of Data Ecosystem	A study that presents a method, tool, program, or application used to create, maintain, or support Data Ecosystems
(d) Research method facet (adapted from [59])	
Case study	Study of a single phenomenon (e.g., an application, a technology, a decision) in an organization over a logical time frame
Research survey	Research that uses predefined and structured questionnaires to capture data from individuals. Normally, the questionnaires are mailed. (Now, fax and electronic means are also used.)
Interview	Research in which information is obtained by asking respondents questions directly. The questions may be loosely defined, and the responses may be open-ended
Field study	Study of single or multiple and related processes/phenomena in single or multiple organizations
Qualitative research	Qualitative research methods are designed to help understand people and the social and cultural contexts within which they live. These methods include ethnography, action research, case research, interpretive studies, and examination of documents and texts
Literature analysis	Research that critiques, analyzes, and extends existing literature and attempts to build new groundwork; e.g., it includes meta-analysis
Proof of concept	A realization of a certain method or idea in order to demonstrate its feasibility
Design science research	A research that develops new artifacts and improves the effectiveness and efficiency of the existing artifacts in the context of solving real-world organization problems

sources for the computer science community and therefore prone to containing the majority of important studies. We also looked for related studies referenced from studies already in the pool, in order to decrease the risk of missing relevant studies.

Data extraction Data extraction could be biased by the personal opinions of researchers executing the process. To mitigate this threat, we defined and documented a strict protocol for selecting studies in addition to which we reviewed. Moreover, we dedicated adequate time to reviewing the studies (which were reviewed a number of times) and addressing the conflicts alongside the issue of missing data.

Moreover, in comparison with other systematic mappings, our study used few researchers. However, the main reason for having a higher number of researchers is to share/reduce the efforts during the filtering, selection, and analysis phases. Some studies deal with thousands of studies, as a result of which considerable effort is required. So, the divide-and-conquer

Table 5 List of selected studies

Study ID	References	Study ID	References
S01	[70]	S16	[87]
S02	[50]	S17	[47]
S03	[82]	S18	[41]
S04	[51]	S19	[9]
S05	[66]	S20	[6]
S06	[26]	S21	[43]
S07	[84]	S22	[33]
S08	[20]	S23	[81]
S09	[88]	S24	[89]
S10	[19]	S25	[30]
S11	[77]	S26	[29]
S12	[39]	S27	[18]
S13	[40]	S28	[80]
S14	[67]	S29	[24]
S15	[76]	–	–

strategy is very useful for these cases. However, in this study, even the amount of studies initially discovered by the automatic search can be dealt with by three researchers in a short period of time. Furthermore, in order to reduce a possible bias from researchers, we specified some mitigation procedures, such as holding a meeting to solve all the conflicts by consensus. Moreover, when we had any doubts about the relevance of a study, during the filtering and selection phases, we always postponed a decision on whether to include it until the analysis phase. In particular, at least four studies were selected by virtue of this procedure.

3 Systematic mapping results

In this section we analyze the literature and the results of the systematic mapping study. From an initial sample of 250 studies, we identified 29 primary studies which fully met our research questions. Table 5 presents the complete list of references of these studies.

3.1 What has been the evolution in the number of Data Ecosystems studies over the years?

We analyzed the evolution in the number of studies related to Data Ecosystems published over the recent years (cf. Fig. 3). The first Data Ecosystem studies were published by Tim Davis [S03] and Ding et al. [S08] in 2011. The former study claims that data initiatives require an ecosystem that consists of open data infrastructures and standards in order to actively encourage the use of open data as well as the development of new technologies. Such a Data Ecosystem involves mobilizing a wide range of technical, social and political resources, and the need for interventions beyond data supply to support the coordination of activity around datasets. The latter study presents a platform called TWC LOGD Portal, which encompasses a model infrastructure that supports linked open government data production and consumption. By using this platform, a community of actors may interact to form an ecosystem. These studies neither create a theory, nor define Data Ecosystems. However, they

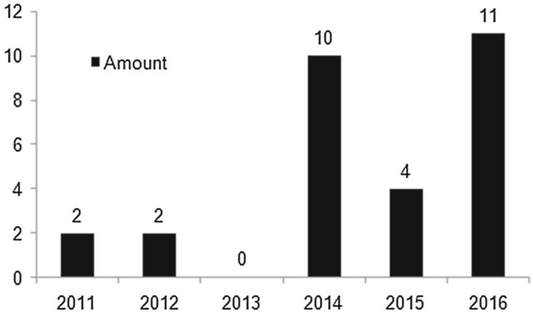


Fig. 3 Distribution of the number of studies published regarding Data Ecosystems between 2011 and 2016. *Source*: the Authors

Table 6 Research focus classification

Research focus	Studies
Full	S01, S04, S06, S07, S09, S08, S10, S11, S15, S16, S18, S19, S20, S21, S22, S23, S25, S26, S27, S29
Partial	S02, S03, S05, S12, S13, S14, S17, S24, S28

provide the first insights into the field. It is important to remark that the most cited oldest references in the literature are the studies [62,63].

We consider that a total of 29 articles on such an important topic is a very small number. However, the number of studies might have increased because we did not consider studies that were published in 2017. Moreover, with the exception of 2013 and 2015, the average number of publications has been growing since the first study was published in 2011. The growth in the number of studies published has been strongly influenced by both the Open (Government) Data movement and the Big Data market.

3.2 What level of focus (full and partial) do the studies dedicate to Data Ecosystem research?

Table 6 relates the studies with the corresponding focus categories (full or partial). For this, we used the classification scheme presented in Table 3. The data collected indicate that over 68% of the works (20 works) fully focus on Data Ecosystem aspects. With respect to remaining 9 studies, they are of partial pertinence to Data Ecosystems research. For instance, S24 and S28 focus on the study of business models and how to articulate capabilities required for open data processes and the creation of value. However, neither of these nine studies define or make reference to a definition for Data Ecosystems, nor do they even describe what ecosystems consist of. Other examples are S12 and S13. Despite these works recognizing the necessity for the systematic development of Data Ecosystems, they focus on the analysis of open data initiatives of the Russian Federation and Uruguay, respectively. No further analysis of the conceptual definitions is presented.

3.3 Who are the main publishers of Data Ecosystem studies?

Table 7 lists the conferences and journals which have published the studies selected. Over 55% (16 studies) of the studies were published in journals. Thirteen studies were published

Table 7 Publication venues for Data Ecosystems studies

Venue type	Venue name	Studies
Conference	Conference on Intelligence in Next Generation Networks	S04
Conference	Hawaii International Conference on System Sciences	S01
Conference	International Conference for E-Democracy and Open Government	S25
Conference	International Conference on Digital Government Research	S02, S03, S24
Conference	International Conference on Enterprise Information Systems	S29
Conference	International Conference on Knowledge Management and Information Sharing	S13
Conference	International Conference on Theory and Practice of Electronic Governance	S12, S17, S18
Conference	International Conference on Web Intelligence	S20
Conference	Working Conference on Virtual Enterprises	S28
Journal	Future Internet	S26
Journal	Government Information Quarterly	S10
Journal	IEEE Access	S22
Journal	IEEE Software	S21
Journal	Information Polity	S09, S16
Journal	Information Systems Frontiers	S23
Journal	Information Technology for Development	S15
Journal	International Journal of Design & Nature and Ecodynamics	S07
Journal	International Journal of Digital Earth	S11
Journal	International Journal of Software Engineering and Its Applications	S06
Journal	Journal of Web Semantics	S08
Journal	Nature	S19
Journal	Open Data Research	S27
Journal	Telecommunications Policy	S05
Journal	Telematics and Informatics	S14

in conferences and workshops. Twenty-one studies are full papers, and 8 are short papers. With the exception of DGO[1] and ICEGOV[2] conferences and the Information Polity journal, just one study was published in the remainder of the conferences and journals. Moreover, 6 of the venues uphold the theme open data or Digital Government research.

3.4 Which individuals and organizations are the most active in Data Ecosystem research?

In the 29 relevant studies, 80 distinct co-authors were found. The most active researchers are Fatemeh Ahmadi Zeleti, Adegboyega Ojo, Anneke Zuiderwijk, Marijn Janssen, each of whom was a co-author of 3 studies; Dong-Hee Shin, Geerten van de Kaa, and Kostas Poulis co-authored 2 studies each. The other 73 researchers co-authored only one study. In addition, we found 35 distinct organizations (universities, research institutions, and companies) located in 17 different countries. The Sungkyunkwan University, University of Albany, Insight Centre for Data Analytics (Dublin), and Delft University of Technology were the

[1] International Digital Government Research Conference.

[2] International Conference on Theory and Practice of Electronic Governance.

Table 8 Research themes and topics addressed in the studies

Research theme and topics	Studies
Functions and features of Data Ecosystems	
Value creation and capabilities	S01, S03, S09, S23, S24, S28
Innovation	S16
Requirements and critical factors	S05, S22, S27
Business models and economics	S06, S17, S19, S20, S21, S22, S29
Describing and modeling Data Ecosystems	
Describe an ecosystems initiative	S07, S12, S18, S25
Modeling and taxonomies	S10, S16, S17, S29
Structuring and shaping of Data Ecosystems	
Roles and relationships	S02, S10, S16, S17, S21, S22, S29
Data Ecosystem architecture	S04
Data Ecosystem solutions	
Tool	S08
Method	S11, S12, S18
Platform architecture	S29
Analysis and evaluation of Data Ecosystems	
Report and analysis	S13, S14, S15, S19, S25, S26

most active organizations with 2, 3, 4, and 4 studies, respectively. As a final note, the most active countries were Finland, Ireland, Netherlands, and the USA with 3, 4, 4, and 7 studies.

3.5 Which topics or themes have been addressed by Data Ecosystem studies?

Table 8 lists the research topics or themes addressed by Data Ecosystem studies. The large majority of them seek to theorize the Data Ecosystem field. Seventeen of the studies focus on describing functions and features of Data Ecosystems. For instance, we found 7 studies dealing with value creation and co-innovation and 7 studies on the business aspects of Data Ecosystems. We also identified a small group of studies (4 studies) proposing models and visual representations for Data Ecosystems. There are also 4 studies documenting a Data Ecosystem experience. Such descriptions do not necessarily reflect key features and mechanisms of ecosystems, but are essential in order to understand how ecosystems function [28].

Moreover, we found 8 studies that explain the roles of actors in Data Ecosystems and actors' relationships. These studies aim to investigate what categories of actors there are, what the actors' duties and activities are, how many actors there are and with what frequency they appear, and what roles they take on. Later, we found few publications that explicitly propose a solution to support some operational or managerial aspect of a Data Ecosystem, such as a platform to ease the interaction among actors [S08] or a guideline to create an ecosystem [S11][S12]. The lack of studies proposing solutions is another sign that Data Ecosystem research is in its initial stages. There are also 6 studies that analyze and evaluate Data Ecosystem experiences. These studies have some practical goals, e.g., to identify or analyze existing problems, or to make an empirical evaluation of Data Ecosystem experiences.

Table 9 Types of research contributions of Data Ecosystem studies

Contribution type	Studies
Tool	S08, S20
Method	S09, S11, S18
Analysis	S14, S15, S19, S26
Empirical model	S01, S05, S10, S25
Report	S07, S12, S13, S18
Descriptive model	S02, S03, S04, S06, S10, S16, S17, S20, S21, S22, S23, S24, S27, S28, S29

3.6 What types of contributions are reported by the studies?

Table 9 relates the contribution categories of the reviewed works. The studies were classified using the classification scheme presented in Table 3. The data collected indicate that the large majority of the studies (15 of the studies) contributes to the body of knowledge with some descriptive or qualitative models. S02, S06, S17, and S22 present and analyze potential business models to define how an ecosystem creates and delivers value to the actors. S03 and S23 propose a model based on the dynamic capability theory for open data driven organizations to support the management of capabilities, processes, and resources that can be used to implement value-creating strategies. S04 presents an architectural model based on a user-centric view, in which the data users have an active role in controlling the use of data and sharing data with each other. S10 and S16 propose a more detailed architectural model to assist in planning and designing Data Ecosystems. Both studies build such a model by considering what the essential components of an ideal Data Ecosystem are. They address data policy and strategy, data publication and use, feedback and communication, the generation of benefits, and advocacy and interaction among actors.

There are also 4 studies that put forward an empirical model produced from observed or collected data. For instance, in [S05][S10], quantitative data based on a survey method were collected, and so too were qualitative data based on an interpretive method in order to understand Big Data Ecosystems and how they are perceived by different stakeholders. This model also aims to study the potential value and challenges, and the implications for Big Data Ecosystem actors. S25 conceptualizes an Open Data Ecosystem model by analyzing the UK's open data initiatives. The conceptualization is based on a review of some relevant studies in open data literature and Business Ecosystem theories as well as in-depth interviews with stakeholders.

There are 4 studies that report or document a Data Ecosystem experience (S07, S12, S13, and S18). While S07 documents the Data Ecosystem created by the University of Alicante, studies S12, S13 and S18 document the Open Data Ecosystem experiences of Uruguay, Russia, and Ireland. Only 3 studies (S14, S15, and S19) analyze a single Data Ecosystem case. S14 analyzes the Big Data Ecosystem of South Korea from a socio-technical perspective, which includes social dynamics, political discourse, and technological issues. S15 analyzes the roles of data intermediaries in the South African Data Ecosystems. S19 analyzes how to sustain Big Data Ecosystems about biomedical data.

With regard to the studies that propose a more concrete solution (e.g., tool or method), two studies propose a tool and 3 studies propose a method. S08 presents a platform for managing an ecosystem and supporting the interaction of actors. S20 presents a set of cloud-based services, called Demand and Supply as a Service (DSAAS) model, for aiding value creation

Table 10 Research methods employed by the studies

Research method	Studies
Literature review	S02, S10, S16, S26, S28
Case study	S01, S04, S10, S11, S12, S15, S23, S27
Survey	S05, S09, S13, S14, S15, S24
Structured interviews	S01, S02, S03, S05, S09, S11, S14, S15, S22, S24
Proof of concept	S06, S08, S16, S20, S24
Qualitative research	S02, S03, S09, S14, S17, S18, S22, S24, S25, S26
Field study	S18

by examining data usage. S09 presents guidelines to support the creation of value from Open Data. S18 also presents a guideline based on the Irish Experience for creating Open Data Ecosystems. For a different purpose, S11 presents a framework to assess the success of an Open Data Ecosystem.

3.7 What research methods have been used in Data Ecosystem studies?

We classified the research method of each study using the classification adapted from [59]. As presented in Table 10, the selected works used different types of research methods: case studies, qualitative research, proof of concept, literature analysis, field study, design science research, survey research, and interviews. It is important to remark that the last two methods are instruments commonly applied in qualitative research in order to collect research data. We did not find studies that conducted controlled experiments, quasi-experiments, and ethnography. These findings are explained by the nature and the current momentum of the Data Ecosystem field.

Qualitative research is represented by over 37% (11 studies) of the studies. Ten studies (over 76%) conducted interviews to collect data, and five studies (17%) applied surveys to collect data. Case studies are the second most common method. These were identified in 8 studies. However, in most instances, the research design of the case study is underspecified, if not addressed at all. Several studies developed a proof of concept to evaluate the feasibility of the solution proposed. But they did not use a formal method to evaluate the feasibility. Furthermore, only one study (i.e., S23) adopted a kind of action research method (i.e., design science research). Five studies focus on an analysis of the literature.

Finally, there are some studies (S07, S19, S21 and S28) not employing any particular research method. Most of them only present some idea or solution, but the research method used to construct or validate the object of study is unclear.

3.8 How is the term "Data Ecosystem" defined?

As briefly discussed in Introduction, there is little agreement about nomenclature and the definition of Data Ecosystems. Although a a discussion in greater depth on terminology is beyond the scope of this paper, in order to be able to analyze the Data Ecosystem field as well as to guide the study process, we should first review some of the existing definitions. Hence, an important goal of this paper is to provide an overview of how the research community defines the term Data Ecosystem.

Table 11 Papers classified according to whether or not they offer an explicit definition for a Data Ecosystem

Definition	Studies
Available	S01, S04, S05, S07, S08, S09, S10, S13, S14, S15, S16, S18, S20, S22, S26, S29
Not available	S02, S03, S06, S11, S12, S17, S19, S21, S23, S24, S25, S27, S28

Table 12 Previous papers most cited by the studies selected

Most referenced papers	Referencing studies
Pollock [63]	S10, S16, S17, S18, S25, S26
Ubaldi [75]	S01, S07, S10, S11, S13, S16, S23, S28
Heimstadt et al. [S25]	S07, S10, S18, S16
Harrison et al. [S26]	S10, S15, S16, S18, S25

Table 11 shows the studies split into two different groupings, namely those that include a definition for a Data Ecosystem and those that do not. Our first finding is that there are a large number of studies (13 studies, over 44%) that do not define the term Data Ecosystem. However, some of these studies make reference to previous studies that had done so. In most cases, this happens because these studies only partially focus on Data Ecosystem research. In some cases, the authors consider a previous study (written by themselves or other researchers) as the basis for the background and definitions needed for what was then their most recent study.

With regard to the studies that provide a definition, 7 of them provide a new definition (i.e., with their own words) for Data Ecosystems. The other 10 studies define the field by using one or more definitions from the existing literature. Furthermore, most of the referenced works (cf. Table 12) do not provide a formal definition. They only conceptualize aspects about Data Ecosystems. For instance, instead of providing a definition for Data Ecosystems, S25 and [75] identify a set of structural Data Ecosystem properties (i.e., circular flow of resources, sustainability, demand that encourages supply, and dependence developing between actors). These aspects are used to construct definitions by other works.

Pollock [63] provides the oldest definition cited in the selected works. According to him, "An ecosystem has data cycles, in which intermediate consumers of data such as builders of apps and data wranglers may share back their cleaned, integrated, and packaged data into the ecosystem in a reusable way. This cleaned and integrated data is often more valuable than the original source." This definition emphasizes the need for data cycles in order to create Data Ecosystems. Besides the cycle, Pollock's vision requires that actors should play roles, such as infomediaries and consumers.

Harrison et al. [S02] envision the idea of a Government Ecosystem, which is a kind of Open Data Ecosystem. According to them, "A Government Ecosystem envisions government organizations as central actors, taking the initiative within networked systems organized to achieve specific goals related to innovation and good government." They also complement this definition by stating that the "ecosystem metaphor [is] often used by policy makers, scholars, and technology gurus to convey a sense of the interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that must be created in technology-enabled, information-intensive social systems, among them, open government." Like Pollock [63], Harrison et al. advocate that roles should be defined, but they also emphasize the idea of a keystone role that controls and coordinates the ecosystem. Moreover, they also recognize a set of contextual factors (e.g., social aspect) as a key element of Data Ecosystem. Furthermore, Harrison et al. pair the ecosystem metaphor with the concept of multiple and varying interrelationships between providers, users, data, material infrastructure, and institutions.

A noteworthy example is study S25 which is referenced by 4 studies. Heimstadt et al. [S25] conceptualize Data Ecosystems by identifying a set of structural properties: circular flow of resources, sustainability, demand that encourages supply, and dependence developing between suppliers, intermediaries, and users.

Other definitions are provided by S07, S08, and S09. According to S07, a Data Ecosystem "is made up of many actors and small organizational structures that should recognize data like the raw material that is in a cycle and is capable of feeding the ecosystem, providing benefits to all parties." This Data Ecosystem view also advocates for a cycle as well as pointing to there being multiple actors, each with their own expectations.

A different perspective is presented in S09, which defines Data Ecosystems as "all activities for releasing and publishing data on the Internet, where data users can conduct activities such as searching, finding, evaluating and viewing data and their related licenses, cleansing, analyzing, enriching, combining, linking and visualizing data and interpreting and discussing data and providing feedback to the data provider and other stakeholders." Similarly, S08 defines Data Ecosystem as "a data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mashup, and consume data in connection with online tools, services, and societies." Both definitions present the idea of activities that would develop some value or benefit to actors who use data. These activities can be assigned to specific roles that will be performed by actors (i.e., stakeholders).

Some shared concepts stand out in the definitions for Data Ecosystems: (1) actors and roles, (2) relationships, and (3) resources. Combining the definitions above with the four elements identified, this suggests that Data Ecosystems consist of a loose set of interacting actors that directly or indirectly consume, produce, or provide data and other related resources (e.g., software, services, and infrastructure). Each actor performs one or more roles and is connected to other actors through relationships, in such a way that actors by collaborating and competing with each other promote Data Ecosystems.

As a matter of fact, despite Data Ecosystem being a term that has been used by a number of studies, the meaning of Data Ecosystem is still very much under construction. It seems that to do so, a rich framework is required to support practitioners and researchers so that they may acquire a much fuller understanding of the current state and potential future of Data Ecosystems.

3.9 What theoretical foundations are adopted by Data Ecosystems studies?

We looked for references to established theories in order to help either to define the term Data Ecosystem, or to develop the background for the research. Similarly to [28], we also looked at either explicit well-established theories, such as socio-technical theory, or wider concepts, such as value creation and co-innovation (Table 13).

Table 13 lists the most common theories used by the authors of the studies selected. Socio-technical systems theory [15] provides a base for 6 of the studies. This theory refers to the systematic integration of social and technical aspects of an organization or society as a whole [21]. The basic idea is that the interaction between social and technical factors influences the outcomes [21] of a process, a system, or an organization. Value chain theory [14] is the second most common theory, identified in 4 studies. The theoretical studies about other

Table 13 Theoretical foundations adopted by Data Ecosystems studies

Theory	Studies
Resource dependency theory [61]	S03, S24
Socio-technical systems theory [15]	S05, S10, S14, S15, S18, S27
Normalization process theory [72]	S05
Information Polity [31]	S15
Actor–network theory [53]	S15
Value chain theory [14]	S19, S20, S22, S28
Dynamic capability theory [16]	S03
Digital innovation [79]	S01
Natural Ecosystems [78]	S25
Business Ecosystems [52]	S18, S22, S25, S29
Software Ecosystems [35]	S16, S29
Digital Ecosystems [54]	S25

categories of ecosystems were used to assist in the conceptualization of the field. In particular, Business Ecosystems studies were referenced by 4 of the studies selected. [52] defines a Business Ecosystem as “an economic community supported by a foundation of interacting organizations and individuals,” which includes customers, producers, competitors, and other stakeholders. The key elements to a Business Ecosystem are the keystone species, which are companies that display leadership and exert a strong influence on co-evolutionary processes [52].

Besides the term Business Ecosystem, it is important to note that the term Data Ecosystem spans ideas that are borrowed from other approaches and for these the terminology varies. For instance, Data Ecosystems are inspired by the notion of Biological Ecosystems, which, in particular, denote a natural unit consisting of all plants, animals, and microorganisms in an area functioning together with all of the non-living physical resources of the environment [78]. Moreover, a Data Ecosystem among other elements includes systems, databases, workflows, people, the market, government, and an infrastructure. Taking into account these elements suggests that a Data Ecosystem needs to combine components from different ecosystems.

The creation of a Data Ecosystem is also prompted by there being a Digital Ecosystem and a Software Ecosystem. According to [54], Digital Business Ecosystems “provide an open source distributed environment, where software components, services, applications and also business models are regarded as digital species that can interact with each other, reproduce and evolve according to laws of market selection.” The term Software Ecosystems is also recent. This refers to networked organizations or individuals, who base their relations on developing, commercializing, and using a central software technology [27,36]. All of these ecosystems break down the internal boundaries of organizations’ production lines by allowing contributions from partners, vendors, and other external parties.

A Data Ecosystem can be viewed as another instance of a Business Ecosystem, a Digital Business Ecosystem, or a Software Ecosystem. Despite sharing network and co-evolution characteristics, Data Ecosystems differ from previous ecosystems. Unlike other ecosystems, Data Ecosystems do not rely on an explicit common platform in which different actors can collaborate. The common platform is actually the wide collections of data exchanged by the actors. In particular, the data do not necessarily need to be provided by a single actor. The lack of a common platform creates a more diffused supply–demand network. Another difference is related to how products traded between the actors are perceived. In Business Ecosystems, business operations and actors per se are the products [48]. In Software Ecosystems, the products are software components or services. In Data Ecosystems, the product is data.

Actually, the boundaries between different kinds of ecosystems are difficult to define [S16]. For instance, Data Ecosystems may entail Software Ecosystems in relation to the network of actors involved in developing and providing data-related software. Hence, understanding these related ecosystems allows us to use some concepts in the Data Ecosystem world. It can provide us with solutions to the challenge of building up information and knowledge about a Data Ecosystem.

In summary, a heterogeneous theoretical background has been used by the Data Ecosystem research studies. This situation is a consequence of two factors: The field is in its infancy, and different research and industry communities have been investigating the area independently.

3.10 What are the main roles of actors in a Data Ecosystem?

In Data Ecosystems, actors can play several roles. Actors are autonomous entities such as enterprises, institutions, or individuals, which act in the Data Ecosystem in one or more specific roles. The actors are motivated by a set of interests [S16]. Actors involved in the ecosystem usually have a commitment to the ecosystem. They may have an incentive for being active in the ecosystem.

A role is a function played by an actor in the ecosystem. It is related to a position and a set of duties. Several roles can be identified in Data Ecosystems. Typically, at least the roles of a data user and a data producer are identified in contemporary Data Ecosystems. However, there are several additional roles, each of which undertakes different duties. Moreover, it is possible to find two different roles sharing the same duty. For instance, both data intermediaries and data providers provide data. Furthermore, Data Ecosystem characteristics strongly determine the need for a particular role. In fact, some roles exist only in very specific ecosystems. For example, in Data Ecosystems based on medical/health data, it is common for there to be a role of assessing ethical issues. Section 3.10 discusses some of the roles found in the literature.

Like the definition of a Data Ecosystem, there is little agreement about the actors' roles and their respective duties and activities. According to [45], defining the roles in each ecosystem is essential for researchers and business analysts in order to understand and manage an ecosystem and estimate its success.

Table 14 lists the roles identified in the studies. In total, we identified 13 major roles and a further 22 minor roles (i.e., a specialized role that is responsible for some of the duties and activities of a major role). However, in most studies, both how the roles are defined and how their activities and duties are set out are underspecified or not specified at all. Moreover, the large majority of the studies only list Data Ecosystem actors (i.e., stakeholders) such as citizens, developers, governmental organizations, and private companies, leaving it to the reader to identify and classify their roles.

The role most often identified is data user, which is responsible for directly or indirectly consuming data. Almost all the studies present the concept of a data user role. However, this role is presented with a myriad of names in the studies, such as end users, data consumers, data beneficiaries, etc. Data Users do not necessarily have the ability to consume data directly from data providers. They usually rely on services provided by Re-users, data intermediaries, or service providers. Moreover, data users usually represent the end users of an ecosystem.

The second most highlighted role in Data Ecosystems is data provider, which is responsible for data supply or provision. Almost all the studies (21 studies) present the concept of a data provider role in a very similar way to that of the data user. There are also a few studies

Table 14 Roles of Data Ecosystem actors

Role	Studies
Data provider	S01, S03, S05, S06, S07, S09, S11, S12, S13, S14, S15, S17, S19, S20, S21, S22, S23, S24, S25, S27, S29
Storer	S22
Developer	S22
Aggregator	S22, S29
Harmonizer	S22
Publisher	S22
Register	S22
Data maintainer	S19
Data producer	S02, S04, S08, S16, S19, S26
Data owner	S21
Policies, Laws and Rules Parties	S03, S09, S11, S10, S15
Policy makers	S09, S11
Standardized and regulation parties	S03, S15
Keystone actors	S10, S15, S19, S26
Founder	S19
Service provider	S05, S06, S14, S21, S22, S29
Data-as-a-service providers	S22
Analytics-as-a-service providers	S22
Data service developer	S01, S29
Re-user	S01, S02, S08, S12, S18, S17, S21, S22, S23, S26, S29
Open data-driven organization	S23
Application developer	S08, S22
Interpreter	S21, S22
Data intermediaries	S01, S02, S04, S09, S15, S16, S17, S21, S23, S25, S26
Data brokers	S01, S04, S22
Enablers	S17, S23
Integrators	S17
Data extractor and transformer	S21
Data analyzer	S21
Data visualizer	S21
Data user/data consumer	S01, S02, S03, S04, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, S15, S16, S17, S19, S20, S21, S22, S23, S24, S25, S26, S27, S28, S29
Data curator	S08, S19
Infrastructure provider	S06, S21, S22
Data sponsor	S19, S29
Data consultant	S01, S22

that present minor roles related to the provision of data. For instance, S22 presents different data provider minor roles related to supplying data, such as "Storer to collect and save raw material, a Developer to manage and process raw material, an Aggregator to combine and edit data from different sources, a Harmonizer to standardize and homogenize data from different sources, an Updater to update information, a Publisher to publish the data, and a Register to maintain the administration of data resources."

It is important to note that data providers are not necessarily responsible for data generation. This responsibility may be assigned to another role, called data producer, which is responsible for the capture or generation of data. This role may also compile, aggregate, and package data. In particular, 6 studies identify the data producer role.

Another identified role is a Re-user responsible for adding value to the data to be reused. According to S12, Re-user is responsible for using data to develop applications or services aimed at data users. The Re-user role is identified in 11 studies. There are several studies presenting the Re-user's minor roles. For example, [S22] defines the application developer role (i.e., they use the data as part of the service) and Interpreter (i.e., they interpret data for end users). Another specialization of Re-users is Data-driven Organization [S23], which represents the private companies that use, transform, or invest in Open Data.

The keystone actor role is responsible for driving forces behind the ecosystem as well as providing stability in unstable environments [32]. This role is very common in Software Ecosystems. In S10, the researchers claimed that in Data Ecosystems, keystone actors are responsible for providing most data as well as for promoting the ecosystem. S18 states that this role must be assigned to actors who lead the Open Government Data (OGD) programs. Under this scenario, keystone actors should foster a set of formal directives, rules, and practices to drive a Data Ecosystem. Taking a slightly different view, S15 states that keystone actors are enablers, not necessarily drivers in the ecosystem; they are useful, but they are not essential to the continued functioning of an ecosystem.

The service provider is responsible for producing and supporting software resources such as tools, applications, services, libraries, or other software products [S05] [S06][S14][S21][S22][S29]. Actors that do not have the abilities and resources to perform the data processing themselves can contract service providers [S22]. S22 presents two minor roles (i.e., Data-as-a-service providers and Analytics-as-a-service providers) related to service provision based on the cloud computing paradigm.

Introduced as one of the most important roles in S10, the Policies, Laws and Rules Parties represent the role responsible for creating the rules and policies to encourage and to control the participation of actors [S03][S09][S11][S10][S15]. Typically, this role is performed by Governments, Data Ecosystem Founders, or Standardization Institutions (e.g., Open Knowledge Foundation and W3C).

Other major roles identified in the literature are the Infrastructure Provider, Data Consultant, Data Sponsor, and Data Curator. Infrastructure Provider is the role that supports the activities of other roles. Infrastructure includes the provision of Information and Communication Technologies (ICT) resources or services such as hosting or storage capacity [S06][S21][S22]. A Data Consultant assists other roles to analyze the possibilities of data and also identifies the actor's needs [S22]. A Data Sponsor is responsible for promoting the open data initiative through both public funding programs and private investments [S19][S29]. Finally, Data Curators are responsible for the quality and availability of data [S19][S08].

Even though we have identified several different roles, there is still a myriad of different minor roles in the literature. However, a more in-depth discussion on classifying roles is classification beyond the scope of this paper.

Table 15 Studies that describe a form of Data Ecosystem organizational structure

Data Ecosystem structure	Studies
Keystone-centric	S07, S10, S12, S13, S15, S16, S18, S25
Intermediary-based	S02, S04, S15, S21, S22
Platform-centric	S08, S12, S27, S29
Marketplace-based	S01, S04
Business model-oriented	S06, S19, S20, S21, S22, S23, S24, S28

3.11 How are Data Ecosystems structured and organized?

In a Data Ecosystem, each actor is connected to other actors by a set of interests or business models. The whole network of relationships may follow an organizational structure, ranging from an ad hoc diffuse approach to a keystone-centric approach [28]. An ecosystem organizational structure takes into account the way the actors are connected and the properties of their relationships, such as relationship dependency [49]. Studying the organizational structure of Data Ecosystems is important to understand and to govern the interaction and organization of actors [12]. Table 15 shows the different groupings of Data Ecosystem structure found in the selected studies. We identified 5 different approaches: keystone-centric, data intermediary-based, platform-centric, marketplace-based, and business model-oriented.

In the keystone-centric structure, actors are organized around a keystone actor, which is directly or indirectly responsible for providing much of the data. However, the keystone actor does not have complete control over the other actors. However, the keystone actor does not have complete control over the rest of actors. They can leave (or enter) the ecosystem at any time. Hence, the keystone actor should be responsible for monitoring, evaluating, making decisions, and taking actions [S25]. There is also a specific kind of keystone-centric ecosystem performed by government administrations or public institutions which has been emerging from open data movements [S10][S12][S13][S18][S25].

The intermediary-based structure depends on the presence of data intermediaries in order to generate value from data. As mentioned in Sect. 3.10, a data intermediary is a role that facilitates the use of data for other actors. Therefore, in a data intermediary-based structure, data providers and data users (i.e., the two extremes of a supply chain) are organized around the intermediaries. Some studies recognize the critical role that intermediaries can play in ecosystems [S15][S22][S21]. They also state that data intermediaries undertake a wide range of functions.

In the platform-centric structure, a platform defines the organizational structure of a Data Ecosystem. According to S10, the platform provides infrastructure and services to support both the provision and consumption of data. S10 and S08 emphasize that the costs of providing data are reduced when the data are released via the platform. The platform can also mitigate interoperability and usability problems. Open data catalog tools (e.g., CKAN[3]) are common, but limited examples of platforms are used to create Data Ecosystems [S29]. According to S29 and S03, Data Ecosystems should show not only data, but also services. They need a platform that provides core services and allows developers to create new services. This platform can be based on a cloud computing infrastructure, for example [S22].

In marketplace-based structures, marketplaces provide the required infrastructures, business models, rules, and services for transactions of data and software between actors [S01]. In

[3] https://ckan.org/.

Table 16 Studies that mention Data Ecosystem domains

Ecosystem domain	Studies
Scientific	S19, S21, S25, S29
Government	S02, S03, S07, S08, S10, S12, S11, S13, S15, S16, S18, S22, S25, S26, S27
Industry	S01, S04, S06, S05, S09, S11, S14, S17, S20, S21, S22, S23, S24, S28, S29

general, marketplaces encompass a technical platform with the capacity to link data providers and data users. They also enable the sale of data, services, and applications. In this sense, S22 states that there are several pricing models suitable for data-related businesses. According to S22, "services and applications can be priced commonly based on features or performance, or the customer is charged a predefined price for customer-tailored services and applications usage."

Despite not defining how the actors should be organized, some studies present business models, which describe the rationale of how an actor creates, delivers, and captures value. In particular, value refers to any benefit that an actor obtains from the Data Ecosystem, such as satisfaction, utility, problem solving, or revenue. Common business models are business-to-business, business-to-consumer, and consumer-to-consumer ecosystems [S06]. According to [12], the business model is important in reasoning the cost, revenue, and/or sustainability of the Data Ecosystem. While S06, S19, and S21 did not present a well-established business model strategy, S20, S22, S23 and S28 and S24 are based on the value chain theory and resource dependency theory, respectively.

3.12 What are the main domains in which Data Ecosystems have been and are being developed?

A Data Ecosystem domain refers to the setting or environment where an ecosystem emerges. It may determine the associated organizational structure and rules. In summary, the ecosystem domains found in our study can be classified into 3 categories: scientific, government and industry. The last category includes all the ecosystems with blurred boundaries, i.e., the ecosystems that are difficult to label by a domain category. Table 16 classifies the studies according to the ecosystem domain presented.

In Scientific Data Ecosystems, actors are concerned with sharing scientific data in academic or scientific communities. A significant amount of the data is open, but there are closed data. The data are usually used to develop new research, validate experiments, and replicate research [S21]. Actors that belong to these ecosystems are from a specific academic area, university, or research institution, and generally, there is one keystone managing the ecosystem.

Mostly, Government Data Ecosystems are based on Open Government Data initiatives and focus on establishing means to engage various categories of actors and on promoting the usage and publication of open government data. Normally, data providers are public agencies and government units, while data users are citizens interested in accountability and transparency, domain specialists that analyze specific issues (e.g., traffic engineers or criminologists), or even companies that want to develop products or enhance the offer of services. S12 and S18 document experiences and lessons learned over the years in their open government Data Ecosystems, such as formalizing a national policy on open data, adopting the regulations required to promote and support the process of making data open, identifying challenges, and presenting mitigation measures.

Government Data Ecosystems can be divided into multiple smaller ecosystems, representing several government levels. For example, S13 defines the coordination of Open Data Ecosystems as a three-level schema: central level, regional level, and municipal level. At the central level, rules and solutions are developed; regional-level governments create local legislation and programs; and the various municipal authorities, from the local level, follow their respective local programs. S07 presents an Open Data Ecosystem covering only one public institution, which has an Open Data Portal that allows the development of the best practices to create Open Data Ecosystems so that universities can foster the reuse of public sector information.

In the industry domain, there are not usually any keystone actors who coordinate the ecosystem. In Big Data Ecosystems (i.e., a notable example of the industry domain) there are several providers, which supply data, services, and infrastructures for consumers or other kinds of actors [S06][S14]. In these ecosystems there is a greater pursuit of innovation, new businesses, and value creation. For instance, S22 claims that actors aim to use data for enhancing products, creating more services and applications, thereby turning open data into a tool to improve competition.

Each ecosystem domain may have thematic sub-domains such as biomedical data, public transportation, marine, personal data, international aid projects and educational data. In the education domain, an Open Data Portal [S07] was proposed at the University of Alicante that allows the development of best practices to implement Open Data Ecosystems for universities. A Personal Data Ecosystem is proposed by S04 as a user-centric ecosystem where people have an active role in controlling the use and sharing of data about themselves. There is a "Bank of Individuals' Data," which is an intermediary role that provides personal data storage and services. The Bank of Individuals' Data enables data users to exploit their personal data and create new business opportunities for all the producers and consumers of personal data.

3.13 What is currently known about the benefits and limitations of Data Ecosystems?

Table 17 shows the benefits expected in establishing/developing a Data Ecosystem according to the selected studies. The most cited benefits are related to enabling or improving aspects of political and social life, such as improvements in the quality of life and social trust, economic growth, the support of policy-making processes, and enhancing citizen services [S01] [S10]. The second most cited benefits are related to economic aspects, such as creating new business opportunities by using data and data services, and enabling innovation and value creation [S06].

Another benefit expected is the ease of consumption and production of data by using Data Ecosystems. Actors that do not have the abilities and resources to consume or provide data can contract service providers or data intermediaries. For instance, S22 presents several specialized roles related to data provision. Another example is S21 which presents the roles of Data Analyzer and Data Visualizer to help data users in the data consumption process. S12 and S15 emphasize that Data Ecosystems also help to promote the interoperability of data and services, thereby aiding the reuse of data and data transparency.

Another great benefit expected is to prompt actors to interact and participate. According to S13, when actors communicate with each other and interact, this contributes to establishing a Data Ecosystem. A well-maintained network with internal and external actors working collectively increases competitive advantages [S24]. As stated in S22, the communication between actors facilitates the delivery of services and sharing of knowledge and also enables

Table 17 Benefits expected from Data Ecosystems

Benefits	Studies
Improvements in political and social aspects	S01, S02, S07, S08, S09, S10, S11, S12, S13, S15, S16, S17, S18, S23, S24, S26, S27, S28, S29
Improvements in economic aspects	S02, S03, S06, S09, S10, S11, S13, S12, S14, S17, S18, S19, S20, S22, S23, S25, S26, S27, S29
Ease of data consumption and production	S02, S04, S05, S06, S08, S10, S11, S12, S22, S24, S25, S27
Communication and interaction between actors	S01, S02, S03, S13, S17, S18, S20, S21, S24, S27
Improvements in the quality of data and services.	S01, S03, S05, S08, S09, S11, S12, S15, S22, S24, S29

several types of partnership and cooperation. Actor engagement and interaction can also be improved by holding events such as hackathons, seminars, conferences, and competitions [S13].

And lastly, Data Ecosystems also contribute to the delivery of better data and services (due to feedback from the actors). According to [S29], data providers and data intermediaries benefit from ideas and feedback about their own processes received as a result of transactions with other actors. Based on relevant feedback, it is possible to improve the ecosystem as a whole by analyzing which applications and services should be continued, revised, or abandoned in favor of alternatives [S29]. In particular, attention must be paid to the quality of the data. For most actors, the data consumption problem is less a matter of data volume than the quality of data [S05]. Actors can give feedback to data providers and thus improve the correctness and quality of the data [S22]. In Data Ecosystems, it is common for there to be rules developed that seek to increase the quality of data [S09].

Table 18 shows barriers and challenges expected that Data Ecosystems identified in the selected studies will face. The most cited barriers and challenges are the lack of technical knowledge and resources to maintain an ecosystem. In S11, the authors express their concern about the long-term sustainability of a Data Ecosystem. According to them, the actors need to generate revenue to cover part of their operating costs. The cost of providing data and the resources needed to make them useful have received relatively little attention [S19]. In particular, the cost of simply keeping the data is usually only a small fraction of the total cost of data management [S19]. From the data consumption perspective, the cost of consuming data is largely related to supporting finding, accessing, and reusing data [S19]. Data Ecosystem actors need to find ways to keep their activities financially and politically viable [S02].

Other barriers pertain to the complexity of activities needed to produce, identify, access, understand, and use data. Even when data have already been created, there are a number of aspects to consider in order to meet the data provision criteria. For instance, the creation of appropriate and sufficient metadata to assist data consumption [S07][S29]. Furthermore, in most cases, data providers do not know what data other actors want and will use. The lack of feedback to effectively engage the actors makes it difficult to know what kind of data is valuable for release. Moreover, the way that data are published also influences possible barriers for using data [85]. For instance, poor data quality [S10][S03], operational changes in the provision of data [S18], and usability problems [S08] are common barriers related to data consumption activities. By using the data consumption perspective, actors should possess several capabilities to find data, collect data, clean and prepare data for consumption, and, finally, consume the data [S16]. The lack of guidelines [S16] and the capabilities required [S03][S23] make it difficult for many actors to use data and to generate value from it.

Table 18 Barriers and challenges for Data Ecosystems

Barriers and challenges	Studies
Lack of technical knowledge and resources to maintain the ecosystem	S01, S02, S03, S04, S05, S07, S08, S09, S10, S11, S13, S15, S16, S17, S18, S19, S20, S21, S22, S23, S26, S27, S29
High complexity of tasks such as data discovery and data consumption	S01, S03, S07, S08, S09, S10, S12, S13, S16, S18, S20, S21, S23, S24, S25, S26, S27, S29
Lack of actor participation and interaction	S01, S09, S10, S12, S11, S13, S14, S15, S23, S24, S25, S26, S29
Lack of organizational structure	S01, S05, S06, S07, S10, S13, S14, S16, S18, S21, S22, S25, S26
Presence of aspects related to privacy, confidentiality, and liability	S04, S06, S10, S14, S18, S26, S28

The lack of participation and interaction between actors affects the ecosystem, causing actors to participate less [S13], and reduces the sharing of data, knowledge, and experiences [S01]. Actor participation and interaction barriers refer to the ease and attractiveness of joining and contributing to a Data Ecosystem [S01]. Underlying factors raising barriers in this area include lack of incentives, lack of ability or time to engage in an ecosystem, costs and competition from other ecosystems [85]. According to S13, it is necessary to provide detailed business, economic, and financial models for the Data Ecosystem to stimulate actors to participate. Furthermore, in order to obtain the best outcomes, it is crucial to raise awareness among the actors, i.e., the actors must understand that value generation calls for a data chain that enriches raw data and thus makes the content valuable.

The lack of organizational structure for Data Ecosystems hinders internal and external participants understanding and, subsequently, implementing and manipulating these ecosystems [S14]. Organizational structure consists of well-defined models for interaction between actors, their roles, specific theories that describe internal and external structures in Data Ecosystems [S05], operation models, policies, and principles [S06]. For instance, in S25 the authors state that the lack of business models that generate value for all the actors causes a one-sided dependency. The same problem is reported by S22.

Additional barriers are derived from concerns for privacy, confidentiality, and liability [S10]. Besides operational activities and technical resources, the requirements for data provision also include privacy and confidentiality protection [S10] [S02] [S28]. Private data are often the data user's own data or information collected about users. In the Web, every day, individuals create personal data that are collected, analyzed, and used by private and public entities. The violation of privacy by opening data and the prospect of being legally liable when data are misused is an important barrier [85]. Existing privacy concerns include the invasion of privacy, imperfections in security, the seriousness of illicit acts, and the misuse of personal data [S14]. In S06 and S04, the authors call for technical standards and guidelines for privacy and security management to be put forward.

4 Data Ecosystem landscape

In Sect. 3, we analyzed the Data Ecosystem literature corpus by positing 14 research questions. Our first finding is the reduced amount of work dedicated to Data Ecosystem. Moreover, a significant number of the analyzed studies have partial pertinence to Data Ecosystems research. The relatively small literature corpus found by this systematic mapping is evidence that the Data Ecosystem area has not been widely explored by many researchers. In fact, while there are a few researchers working in the area, there is no center dedicated to conducting research into Data Ecosystems.

Despite Data Ecosystems being in their infancy, researchers have tried to study them from different perspectives, but the terminology and connectivity between Data Ecosystem concepts are still vague. There are few explanations of what a Data Ecosystem is, how it behaves, what that implies for the design and management of Data Ecosystems, and what factors need to be considered in order for a Data Ecosystem to succeed. These findings indicate that an imprecise notion of Data Ecosystem may be insufficient for advancing this research area.

In this systematic mapping, one of our initial aims was to identify various essential elements and features of Data Ecosystems. The definitions and concepts presented by the studies analyzed assisted characterizing Data Ecosystems. The identified concepts provided some clarity on what is known about Data Ecosystems. Furthermore, we also obtained an overview of real-world Data Ecosystems by mapping which of them were their drivers, how these ecosystems were created, and what the available infrastructure is. Finally, by identifying the actors and their role as well as the Data Ecosystem organizational structure, we obtained an understanding about models and frameworks that can help in understanding the process and activities related to Data Ecosystem management and the generation of value.

In this context, the objective of this section is to provide an overview of Data Ecosystems based on these findings to enable a better understanding of this field.

4.1 Data Ecosystem characterization

Ecosystem metaphor is often used "to convey a sense of the interdependent social systems of individuals, organizations, material infrastructures, and resources that can be created in technology-enabled, information-intensive social systems" [S02 and S16]. This view is similar to that of S14, who argue that Data Ecosystems are a cultural, technological, and social phenomenon based on the interplay of technology, actors, businesses, industries, and governments. Similarly, S10 argues that Data Ecosystem understanding should place the emphasis on an evolving, self-organizing system of feedback and adjustment among actors and processes. S16 emphasizes the multiple and varying interrelationships between data, actors, infrastructure, service, tools, and other different kinds of resources. These connected components are extremely interdependent [S16]. These ecosystem views suggest that an ecosystem is a socio-technical system, the design and analysis of which should be based on a contextual understanding of human interactions that are served by technology, but also on practices, culture and the political and economic context [S15]. Indeed, successful Data Ecosystem initiatives emphasize the importance of considering all components that constitute a socio-technical system [S12][S14][S18][S07].

In general, a Data Ecosystem addresses the multi-aspect nature of data sharing initiatives, encompassing several dimensions such as the social, legal, political, cultural, technological, operational, and economic ones. Such a holistic nature emphasizes that all aspects of an ecosystem are interconnected, influenced by values [S10]. As a matter of fact, Data Ecosystem initiatives in different countries have faced many barriers because of differences in government organization, legislation, and other factors [S13]. Therefore, the design and study of a Data Ecosystem should include seeking to understand the context.

According to S15, in Data Ecosystems, there are at least three contextual conditions that can motivate or constrain the functioning of Data Ecosystem elements. The first of these is the regulatory context, which includes laws, policies, standards, and agreements. This context guides how to specify how the elements of the ecosystem are structured and how they interrelate [S15]. The second is the institutional context in which the actors operate. Each institutional context provides values, rules, and norms that inevitably propel and restrain the behaviors of actors in the ecosystem [S15]. The third is the technological context, which encompass the IT resources, the IT operators, and other enabling technologies that connect and interconnect the Data Ecosystem elements [S15]. The environmental context, such as cultural, social, and economical elements, also exerts an influence on Data Ecosystem initiatives [S18].

With respect to other Data Ecosystem properties, the current literature also conceptualizes that Data Ecosystems are cyclical, sustainable as well as demand-driven and there is dependence between suppliers, intermediaries, and users [S25]. By cyclical, the authors states that any resource should be processed cyclically. Sustainability is understood as the ability to continue a defined behavior indefinitely [58]. Analyzing the UK Data Ecosystem, S25 found "clear signs of a cycle, of sustainability, of demand encouraging supply, and of dependence developing between suppliers, intermediaries, and users." With regard to demand–supply, Data Ecosystems instead of presenting a mutual interdependence often demonstrate more of a one-sided dependency, i.e., data users depend on data producers, or vice versa [S25]. Despite these gaps and drawbacks, the structural properties presented are still desirable, and Data Ecosystem emergence should be fostered by the self-organization aspects rather than the explicit design goals of conventional IT [S14].

4.2 Data Ecosystem elements

As to Data Ecosystem elements, in a literature review, S02 points to three important domains of elements, which are: (i) government policies and practices; (ii) innovators, a combination of technology, business, and government; and (iii) users, civil society, and business. S14 also identifies as key elements of a Data Ecosystem (i) infrastructure, (ii) software and technologies, (iii) service and applications, (iv) standards, (v) users, (vi) social and cultural factors, (vii) government, and (viii) industry.

In order to create and capture value, S23 states that actors must employ emerging sets of capabilities, which are types of skill. Their function allows an activity to be performed or even to improve the productivity of some activities. These capacities can include a range of skills and expertise, such as general knowledge about Data Ecosystem resources (e.g., systems and technologies), technical skills required to manage and to process data (e.g., data integration and data mining techniques), and operational expertise to incorporate data-related resources into existing institutional and business processes [S18].

Besides capabilities, actors also require proper resources in order to provide and consume data. Common examples of resources are datasets, services, tools, financial capital as well as human capital, equipment, materials, and proprietary technology. S24 distinguishes between three categories of resources: human resources, data resources, and IT resources. Human resources refer to individuals who use their capabilities to explore and exploit data. Data resources refer to the static and dynamic data-based assets, such as databases, knowledge bases, or simply datasets. IT resources refer to hardware (e.g., infrastructures, networks, and computers), platforms and applications (software). Actually, actors do not necessarily need to own, manage, or operate the underlying resources, but can consume or contract such resources through other actors, such as service providers or other kinds of intermediary actor. S24 also emphasized that these resources often need to be combined to be able to address an actor's expectation. Not only data are a primary resource, but also various kinds of resources are complementary and are needed so that Data Ecosystems function properly.

S02 emphasized dependency relations in Data Ecosystems. According to these authors, all Data Ecosystem elements are interconnected in a way that when one element is changed, effects can be felt throughout the whole system. In fact, actors affect and are affected by the creation and delivery of resources performed by the other actors [S22]. Moreover, actors have their own interests and benefits which could lead to conflicts. In particular, data consumers are strongly influenced by the decision of a data provider to not publish or update a certain piece of data (anymore), to change the format in which the data are published, to compromise the quality of the data, or to change how it can be used [S03].

4.3 Data Ecosystem creation

Data Ecosystems are initially created by pursuing potential social and economic benefits [S11]. According to S02, Data Ecosystems are also created in part by executives and government administrations that want to achieve potentially beneficial factors [S02]. Moreover, within government, industry or academy, data are created, registered, processed, kept, used, shared, and published for certain purposes [S12]. However, these available data are capable of other uses. In the case of governments, Data Ecosystems are not only driven by financial incentives or rewards. Many government administrations have encouraged society and private institutions to consume government data to create economic and social value [S15].

Similar beliefs across the world have spurred a growing number of private companies seeking to release internal data in order to improve their image and even increase profits by enabling external individuals and other companies to analyze information and come up with valuable findings and service/product innovation [S21][S23]. According to S22, private companies can earn direct profits from data sales in addition to which creating a Data Ecosystem centered on their private data can provide other benefits, such as new partners, new interests in the company's main products, new kinds of business activities, and new customers. For instance, in 2011, the European Commission estimated the economic impact of directly and indirectly using open data in the 27 countries then in the EU to be about $140 billion annually [S21][S24][S23].

In fact, the move toward Data Ecosystems is not only desired, but inevitable. The marriage of political goals for transparency and contemporary ICT tools reduced the cost of information capture, management, and use. These new possibilities contribute to the pressure on private and public institutions to make data and documents available [S02].

Data Ecosystems can be seeded, modeled, developed, managed, i.e., intentionally cultivated, for the purpose of achieving a managerial and policy vision [S02]. In fact, many Data Ecosystems were created by using data sharing programs, which typically comprise a set of formal directives, rules, and practices that apply to all or most actors within a community or institution. Moreover, these Data Ecosystems may be conceived from scratch or by extending the existing infrastructure platforms, such as CKAN[4], which allow the amount of work needed to establish an ecosystem to be reduced.

S10 presents five different Data Ecosystem development approaches, namely: (i) data-oriented approach focusing on the characteristics, quality, and availability of open datasets; (ii) program-oriented approach addressing the purposes and features of a data sharing program

[4] https://ckan.org/.

with emphasis on data release initiatives; (iii) use and user-oriented approach focusing on the factors that influence data use by actors; (iv) scorecard and impact approach addressing a wider array of considerations that influence how and how well Data Ecosystems work; and (v) network-based approach focusing not only on Data Ecosystem elements, but on their dynamic relationships and their influence on the ecosystem performance. All these approaches allow the emergence of a Data Ecosystem to be stimulated. However, an ecosystem does not develop solely through top-down governance. It is crucial to provide role and benefits for all the Data Ecosystem actors, since an ecosystem depends on fruitful interaction between cooperating and competing actors [S25].

4.4 Data Ecosystem infrastructure

A number of infrastructures have been developed in recent years to create Data Ecosystem platforms in order to explore the potential of data, such as Open Data Portals, the European open data infrastructure [5], infrastructures of statistics (e.g., Barometer[6]), data publishing solutions (e.g., CKAN, Junar[7], and Socrata[8]). Another alternative is to use services as building blocks so as to construct platforms for Data Ecosystems.

In general, these infrastructures support or accelerate the access and exchange of resources, mainly data, between the supplying actors and consumer actors. This may impose tailoring or repacking resources to fit to technologies such as Web applications, databases, and APIs so that both computing-skilled actors and domestic actors can navigate, explore, and subsequently gain insights from the Data Ecosystem resources. Moreover, such infrastructures must also promote the interactions between the actors of the Data Ecosystem and they should also enable innovative products and services to be created in large scale by using standard formats and common tools.

In fact, the literature on Data Ecosystem frequently advocates for efforts to establish rich new data infrastructures and standards and to actively encourage the development of Data Ecosystems [S04][S14][S03]. However, providing a simple and reliable infrastructure involves more than the release and aggregation of those datasets. The lack of standardized technologies and well-accepted guidelines may hinder ease of use and reduce productivity. For instance, the myriad heterogeneous solutions imply that actors must deal with different APIs and data access methods, the structural, syntactic, and semantic heterogeneity of data, and the variability of quality levels and the diversity in types of metadata.

Promising alternatives are intermediary roles to simplify and promote the access to resources. For instance, [62] define the intermediary roles Aggregator (to combine and edit data from different sources), Harmonizer (to standardize and homogenize data from different sources), application developer (to utilize the data as part of the service), and Interpreter (to interpret the data). Hence, those actors that do not have the capabilities and resources to discover, access, and process data themselves can delegate these activities to intermediary roles. Intermediary roles also can provide the infrastructure to support other Data Ecosystem activities.

Another option is the creation of marketplaces to facilitate trading and sharing of data, services, and other resources. Since data itself become an asset, they may be traded in marketplaces as a commodity along with goods and services. For example, Microsoft Win-

[5] https://www.eudat.eu/.

[6] https://opendatabarometer.org/.

[7] http://www.junar.com/.

[8] http://www.socrata.com/.

dows Azure Data Marketplace[9], as the name implies, integrates data with its applications. Another example is Bloomberg Marketplace[10] that brings data from a variety of sources and providers, curates, and makes them available to its customers who pay-per-transaction and/or subscription fees. According to [57], digital marketplaces in general provide technological infrastructure, rules, business models, and services for transactions between providers and consumers.

4.5 Data Ecosystem value generation

In Data Ecosystem, actors are required to employ a set of capabilities and resources to catalyze value. According to S17, often the onus is on the consumers to extract value from the available resources. This creates a problem since the average consumer lacks the necessary skills [86]. Due to these barriers, value is not created by a single actor but rather a value chain (i.e., in a network of actors). A value chain is a set of independent value-adding activities that is used to exploit a set of resources. Moreover, a value chain consists of different actors conducting one or more activities (e.g., data provision, data curation, data analysis), and each activity can consist of a number of actions or value creation techniques (e.g., gathering, visualization, service creation). In Data Ecosystems, the minimal value chain consists of data providers, data intermediaries, and data consumers [S25]. As value-adding activities offer different complexities, it is possible that each action can consist of one or more value chains [S20].

The introduction of incentives and rewards can stimulate the flow of resources in a Data Ecosystem [S04][S21]. In fact, the production, provision, and exploitation of Data Ecosystem resources need investments [S14][S21]. S14 states that without money, it becomes very difficult to sustain Data Ecosystem initiatives. However, there is little incentive to invest in resources and capabilities [S22][S10]. The lack of knowledge on the benefits of sharing data and the lack of new operation models are the main impediments that explain why actors, mainly private companies, are not currently motivated to engage in Data Ecosystems [S22]. Therefore, it is important to develop sustainable business models that give them an incentive to keep data up-to-date and accessible and, in addition, to create sustained commercial applications and tools [S21][S22].

Business models support the value proposition for actors in an ecosystem. A number of business models applicable to data were described in the literature [74,83], such as support, subscription, and professional services/consulting business models, which could be applicable both for data providers and for application developers [S22].

Figuring out how to earn revenues (i.e., earn financial profit) from providing data or data-based resources to consumers is a key element of business model design in Data Ecosystems. Services and applications can be priced commonly based on features, cost or pay-per-use go models, for instance [S22]. Data resources can be priced using the subscription model, in which the consumer pays a fixed price for a certain time frame. Another suitable model is the Flickr multiple revenue stream model [74] as it involves collecting subscription fees, charging advertisers for contextual advertising, and receiving sponsorship and revenue-sharing fees from partnerships. However, in general, data are often complex to price, and consumers have many ways to obtain certain types of data without paying [74]. Another alternative is to earn revenues by attempting to extract meaningful and actionable information from it in order to support other business, such as targeted advertising and improvement of services and products.

[9] http://datamarket.azure.com/.

[10] https://www.bloomberg.com/professional/product/market-data/.

4.6 Data Ecosystem management

Although Data Ecosystems have the potential to generate benefits, many ecosystem initiatives are struggling to establish effective management of their resources and actors [63,75]. In fact, while the potential of Data Ecosystems is real, the realization is unsuccessful in many cases [S25][S02] [86]. For instance, there are fundamental concerns related to Data Ecosystem development despite high expectations and financial investment. According to S10, as a consequence of numerous barriers and limitations, the performance of Data Ecosystem initiatives tends to be simplistic. They focus on releasing and promoting short-time contests, such as hackathons and code fests. Even the applications developed for government-sponsored application contests present unsatisfying outcomes. Most applications in such scenarios end up being quickly abandoned. From a long-standing perspective, very few applications resulting from such contests are actually scalable and sustainable [S29].

Establishment of the right Data Ecosystem means the proper coordination of various categories of actors and the provision of business support and stimulation of resources development and usage [S13]. Other essential elements in a successful Data Ecosystem are actor collaboration, integration of scientific, social, and economic information, preservation of ecological processes, and adaptive management [S16]. In fact, Data Ecosystem management is important in order to actively facilitate the effective functioning and accomplishing the goals of ecosystems [S02]. Moreover, if a Data Ecosystem does not have a management structure, it becomes difficult to drive the ecosystem forward and to build and learn from past experiences [S18].

Data Ecosystems will perform well only if they are designed with an appreciation of their full complexity. According to S02, the management of a Data Ecosystem requires sketching some basic topics, that focus on (i) identifying the most active actors that act as essential components of the ecosystem; (ii) analyzing the nature of the transactions that take place between those actors; (iii) recognizing what resources are needed by each actor and how they engage transactions; and (iv) studying the indicators that signal the status of ecosystem activity [S02]. Therefore, these considerations demand a more systemic approach to program planning and designing Data Ecosystems.

However, so far, Data Ecosystem management initiatives are simplistic. Governments around the world have developed programs and policies that aim to promote both the supply and use of public sector data [S02][S13]. Such policies very often focus on ensuring the availability and quality of data resources. These management policies have helped to expand Data Ecosystems and improve the provision of data. However, such policies fail to include other key actors such as data consumers and data intermediaries, who actually demand the supply. Hence, it is crucial to include, from the beginning, the point of view of all the ecosystem's actors. An integrated and collaborative management must ensure that goals included in a Data Ecosystem agenda would meet the needs, rights, and interests of all actors who are part of the ecosystem [S12].

With a different management approach, S14 recommends the intervention of government to trigger or guide Data Ecosystem development and to link these with the government's sector objectives. Among a list of recommendations, S14 advocates governmental investment in a core data infrastructure that will become the foundation for advanced Data Ecosystems. Moreover, governments need to leverage social forces and integrate them into technological arrangements when implementing the ecosystem as a long-term, advanced development strategy [S14]. Similarly, S03 also argues that the role of government should be to provide a simple, reliable and publicly accessible infrastructure that exposes underlying data, with any use of such data left entirely to other actors, either nonprofit or commercial ones. However, S03 states that governments should provide an infrastructure and then get out of the way in order to promote self-organization for the ecosystem.

Assessment and evaluation methods are useful tools to improve the effectiveness of Data Ecosystems. The literature refers to the concept of health of an ecosystem as a means to monitor and assess ecosystem functioning, identify and predict areas for improvement, and evaluate changes in the ecosystem [48]. Until now, alternatives to measure Data Ecosystem health are still naive, as the focus is on relatively simplistic metrics such as number of datasets published, number and percent of existing datasets downloaded, number of datasets scheduled for release, number of APIs, and basic site analytics (e.g., number of page views, downloads, etc.) [S10]. While informative to some extent, these health indicators focus only on the data provider perspective and as a consequence they do not evaluate how useful are the resources provided.

5 Research directions

The purpose of this paper is to find current relevant research on Data Ecosystems as well as to provide an overview of the field. Apart from providing an overview of the Data Ecosystem field, we also identified a number of aspects that are not covered in the literature *corpus*: classification scheme for the role of actors, governance method, engineering solutions, platforms, and so on. Moreover, up until now, there are not many academic papers related to Data Ecosystems. In most cases, they are focused on some specific aspects, such as business models, or a solution that reflects only a small fragment of the whole research area. In the following sections, we discuss some important areas to be considered for developing and evolving the Data Ecosystem field: theory, models, engineering, and solutions.

5.1 Data Ecosystem theory

Data Ecosystems have been driven primarily by the needs of some key actors [S14], while a theoretical basis has been slow to develop. According to Stol and Fitzgerald [73], theories provide a shared understanding that allows different researchers to discuss a topic of study. Theory also allows one to define a context and coverage for research studies. Moreover, theory enables one to make understandings and explanations of how a research object works or how the research object is characterized [73]. In summary, theory has an important function in making knowledge transferable.

However, until now, there is no common agreement on what theories should look like in Data Ecosystems. A piece of evidence for this is the lack of a well-accepted definition for the term Data Ecosystems. Such a definition would provide a conceptual basis for further ecosystem technology development and management. So far, the terminology and definition for Data Ecosystems have shown great variation. As presented in Sect. 3.8, we identified more than 15 different definitions. This diversity imposes a problem for developing a clear understanding of Data Ecosystems. Moreover, the whole functioning of Data Ecosystem is still ambiguous in practice. It is not clear how to design, maintain, or make Data Ecosystems evolve.

Using empirical research is one approach to construct theories by studying and learning how and why things work. Empirical studies take many forms, such as formal experiments, case studies, surveys, and prototyping exercises. Their essence is gathering information on the basis of systematic observation and experiment, rather than deductive logic or mathematics [69]. Therefore, empirical studies can help to understand how practitioners construct, maintain, and evolve Data Ecosystems. Such studies should focus on investigating not only the tools and processes used in Data Ecosystems, but also the social and business processes surrounding them. These studies can be used to yield insights including what are the most efficient business models or which are the most precise metrics for gauging Data Ecosystem effectiveness or sustainability. For instance, the empirical studies presented in S05, S10, and S14 seek to understand Data Ecosystems according to socio-technical frameworks. Although these help to understand some essential ecosystem features and behaviors, it is still necessary to conduct new empirical cases in order to identify additional components or detailed factors within Data Ecosystems as well as to validate the constructed theory in other specific scenarios, such as Private Data Ecosystems.

It is also relevant to consider the role of descriptive studies used to describe a phenomenon of interest and related constructs and relationships [64]. Some authors have argued that descriptive studies can be a valued research method when no explicit theoretical framework is accepted [64]. For instance, S10 and S16 performed descriptive studies for describing the life cycle of provision and consumption of data in Data Ecosystems. While the former focuses on a wide range of social and technical factors that affect the nature and performance of open data sharing programs, the latter focuses on activities related to production, provision, and consumption of data.

We advise that future work should further and more deeply investigate, develop, and relate the content of the theories and fundamental concepts of Data Ecosystem that have been identified so far. Furthermore, we also believe there may be valuable knowledge and inspiration to find in related and more established research fields such as Business Ecosystems and Software Ecosystems. Replicate studies done in Software Ecosystem such as [12,27,42] can help to understand why and how a Data Ecosystem occurs and how it evolves. Further research should be done in order to describe other important Data Ecosystem aspects, such as to understand how relationships are formed and how they are characterized. They should also describe and compare how different factors affect the performance of actors. Finally, the role of keystone actors needs further work to understand how and why this contributes to the design, health, and performance of Data Ecosystems.

5.2 Data Ecosystem models

The term model is generally used to denote an abstract description of a study object (or part of it) (e.g., concepts from the real world, behaviors, or systems) related to a specific point of view [5]. A model should be able to answer questions in place of the actual study object. Models must be precise enough to be subject to automation, and therefore, it is important for them to be written in a well-defined language. Models support a wide spectrum of objectives ranging from facilitating human understanding and supporting process management to automating the execution of some activity [5]. According to Silva [17], models also allow more effective and efficient planning to be undertaken while providing a more suitable view of the system to be developed. They also allow system control to be achieved according to objective criteria.

Data Ecosystem covers a wide range of disciplines including design, orchestration, management, and assessment. For each discipline, models help in understanding the functioning and activities of Data Ecosystems. In summary, models are a kind of blueprint that can be used for running and managing Data Ecosystems. They also may help to define strategic plans for achieving ecosystem's goals, such as value creation and new businesses. Hence, a model tends to provide the means for developing a framework to control and manage an ecosystem.

The models identified in the selected studies were classified as empirical models or descriptive models. In particular, most of the studies present and analyze potential business models to define how an ecosystem creates and delivers value for the actors. Despite being important, these kinds of models cover only a small fragment of how a Data Ecosystem works. In the following subsections, we present other relevant models.

5.2.1 Conceptual models

To the best of our knowledge, a conceptual model for describing a Data Ecosystem and its essential elements has not been proposed yet. In particular, we have found no evidence of a conceptual model that (i) defines the constructors and the construction rules required to describe the core elements of a Data Ecosystem, (ii) can be used as a reference for studies aiming to specialize (or reuse part of) in specific Data Ecosystem cases (e.g., Open Data Ecosystems or Big Data Ecosystems), and (iii) allows interchange or the transformation of models between Data Ecosystem tools.

Therefore, explicit and formal conceptual models are needed on the one hand for enabling efficient management practices, and on the other hand for describing the knowledge about the capabilities and other characteristics of a Data Ecosystem. Moreover, such conceptual models can be used as meta-model language to develop CASE (computer-aided software engineering) tools to help practitioners to construct derived conceptual models that would represent specific Data Ecosystems (e.g., Biomedical Data Ecosystems and Financial Analysis Data Ecosystems).

5.2.2 Modeling languages

In fact, there is also a lack of modeling languages for representing Data Ecosystems at a high level of abstraction. For instance, business actors, or even technical users, may face some difficulties when evaluating a business model or the structural organization of a Data Ecosystem. A new modeling language based on a graphical notation can provide the capability of understanding important aspects and processes of a Data Ecosystem as well as giving actors the ability to communicate these aspects and processes in a standard manner.

An alternative would be to extend an already recognized language, such as BPNM [68] or to create a brand new language based on a flow charting technique. Such an extension does not alter the semantics of the language, but provides some new notations to model specificities of Data Ecosystems. Going further, the creation of a family of modeling languages, each offering a perspective of a Data Ecosystem would provide a simpler and easier way to gather insight into Data Ecosystems. Finally, models proposed for Software Ecosystems (e.g., [10,36]) could be adapted for representing Data Ecosystems.

5.2.3 Assessment models

Assessment models and solutions for validating the health of Data Ecosystems are another gap in the Data Ecosystem literature. These models should provide the means to evaluate the functionality and status of elements in a Data Ecosystem. The health of an ecosystem depends in part on a variety of factors, including the actors and how they act, relationships, policies, and the infrastructure available. In fact, part of defining effectiveness and success for a Data Ecosystem lies in determining and utilizing metrics to measure its health [S02].

However, very few studies set out in detail, analyze, or measure the health of a Data Ecosystem. For instance, we identified only one study addressing this theme (i.e., S11). Therefore, we recommend that future research also involves the proposal of health assessment frameworks to monitor ecosystem activity as well as to identify and predict areas for improving and evaluating changes in the ecosystem. Such further research would define and operationalize a comprehensive set of metrics to determine the health of a Data Ecosystem.

5.3 Data Ecosystem engineering

A Data Ecosystem needs to be organized and to some extent managed. However, since it is a relatively new field, many actors do not know how to effectively manage their Data Ecosystems. Therefore, engineering methods provide guidance on developing or improving processes and systems to meet their users' goals [71]. Engineering disciplines usually provide principles, methods, and tools used to develop, analyze, implement, and verify the systems. For instance, the software engineering discipline provides a set of processes, activities, and tasks that can be adapted according to different software projects [71].

Developing engineering methods for effective Data Ecosystems governance and control is another gap in the Data Ecosystem research. Data Ecosystem functioning depends on the activity and interaction of a set of different actors. Such diffuse performance comes at the expense of decreased control and the resulting increase in challenges associated with planning and maintenance. Therefore, Data Ecosystem engineering methods supply a common structure in the form of well-defined rules, procedures, protocols, and processes to develop, manage, and evolve Data Ecosystems. Engineering methods can also increase the ecosystem's health.

5.3.1 Development methodologies

A Data Ecosystem development methodology is a kind of systematic process that divides Data Ecosystem creation into distinct phases to improve design, management, and evolution. It is often known as the Data Ecosystem life cycle. The current alternatives are the creation of road maps, visions, or long-term strategic planning that allow the actors to plan their activity in the ecosystem and align their business models with the ecosystem agenda. This is the case presented in S04 and S14, which present the efforts of Korean government to coordinate a local Big Data Ecosystem initiative by editing a set of regulations regarding Big Data Ecosystem, such as a general guideline of customer privacy protection and a data regulatory compliance and standard requirements. Other studies, such as S07 and S12, present a road map of the strategy used to create their Data Ecosystems. However, these road maps lack a detailed and systematic process besides which they were designed to address cultural and institutional issues of their countries. Hence, additional work would be needed to formulate guidelines for different Data Ecosystems domains.

Future work that needs to be carried out is to develop methodologies for Data Ecosystem, thus providing some standard way to guide how different ecosystems should be developed. These methodologies should provide the responsible actors with clear guidance on how the Data Ecosystem initiatives should be implemented and how the known challenges and risks should be addressed. If the challenges are not properly tackled, it might prevent the generation of expected benefits. Furthermore, a proper methodology must drive to a clear project with well-defined objectives and actions, as well as clear planning.

5.3.2 Management and coordination frameworks

The management and coordination of Data Ecosystems is another research gap. A management framework can be defined as a set of various methods to discover, model, analyze, and measure the coordination of actors to improve their activities as well as to deliver more benefits. Specifying a management framework faces some fundamental challenges. One of these is the link between the plethora of Data Ecosystem elements and their contribution to value generation. Moreover, a Data Ecosystem management framework must encompass regulations applicable to industry, policies, and quality standards. In short, the distributed and socio-technical nature of Data Ecosystem requires a systematic, holistic approach to management that aligns actors' capabilities and resources to their needs.

Hence, further research should design and evaluate a comprehensive Data Ecosystem management approach to align actors' objectives, adequate governance, and resources. Moreover, we also recommend looking into how actors interact, communicate, and collaborate. Such analysis will allow the drawing up of prediction models, guidelines, and/or best practices to control and improve actor activities or even to make the Data Ecosystem more attractive to its actors.

5.3.3 Governance frameworks

A proper formalization for Data Ecosystem governance is also lacking. [3] defines governance as managerial tools of orchestration actors in Data Ecosystems that have the goal of influencing an ecosystem's health. Governance also encompasses a set of principles to direct the distribution of duties and rights among stakeholders [3]. Thus, governance mechanisms are employed to establish the level of control over a institution or even a community of actors.

Despite there being an extensive body of knowledge on governance in other research fields, further research needs to be conducted on developing governance mechanisms for Data Ecosystems. In addition, there is a need for further understanding of how to implement governance to generate value and benefits as well as of how to incentivate actors to engage and participate in Data Ecosystems. Finally, it is important to provide practical and strategic guidance for Data Ecosystem practitioners to implement governance mechanisms.

5.3.4 Maturity models

A maturity model is a means by which to evaluate the capabilities of an organization with regard to a certain discipline. Maturity models became popular from when the Capability Maturity Model (CMM) was first proposed. According to [22], maturity models are used to compare and evaluate improvements, thus allowing the degree of evolution in certain domains to be measured. In the business environment, they aim to help organizations identify ways to improve the quality of their processes and reduce their execution time, thereby providing them with competitive advantages. In Data Ecosystems, maturity models can be applied to provide clear recommendations on how to drive improvements based on knowledge of the maturity level in which the ecosystem lies.

An alternative to assess maturity in Data Ecosystems is to use objective measures as a guideline. Another example of how to define maturity is to use multiple dimensions (e.g., number of resources, number of actors, average degree of relationships networks) to rate the Data Ecosystem condition. A Data Ecosystem maturity model would follow a thematic approach; for instance, it would measure the degree of management and coordination of a Data Ecosystem, from ad hoc practices to formally defined management process. As a matter of fact, in this literature review, we did not find any maturity models for Data Ecosystems. In the course of this, there is a need for research directions in this area, since innumerable benefits can be gained by using maturity models within the Data Ecosystem.

5.4 Data Ecosystem solutions

There are several technical/non-technical challenges related to using data in a Data Ecosystem, including the complexity of activities needed to identify, understand, and use data, lack of capabilities and technical knowledge among actors [37,85]. Additional challenges encompass problems with the provenance of data, data management, and quality (e.g., validity, completeness, and timeliness), metadata provision, and interoperability, as well as concerns for privacy and confidentiality [37,85]. Generally speaking, the proposal of solutions to address some of these challenges may ease the burden on actors, mainly on data consuming actors, and consequently promote their participation in Data Ecosystems.

Although a deeper debate about promising solutions is beyond the scope of this paper, this study suggests some future lines of research. For instance, there is a great need for data processing, analysis, and integration services [S10]. Data processing provides additional value for data and the integration of diverse data sources may generate value. Data analysis (e.g., mining and extracting) solutions have a high potential in business. Furthermore, cloud-based services are needed in order to provide solutions to scalability, capacity, and interoperability problems when dealing with large volumes of data or costly processing algorithms. In addition, further research should be conducted in order to optimize algorithms to deal with great volumes of data.

With regard to data provision, heterogeneous data, nonstandard APIs, and varying licensing conditions complicate the use of data. A Data on the Web Management System [56] would make it easy to define, create, maintain, manipulate, and share datasets across multiple actors and applications. This system may be seen as a collection of services that allow users to share data. There are research studies that target this area, but nothing concrete. Furthermore, there is also a need to provide more flexible data provision methods for real-time data, such as data streams.

Moreover, there is also a growing interest in refined data/information in order to support predictions and decision-making processes. In addition, several actors have concerns about the quality of data [S09,S10]. The availability and quality of these particular types of resources must be ensured, so that data value creation can be stimulated. A promising solution is to use a well-conceived, efficient curation strategy for data resources and their metadata. Such a curation technique is the continuous process of managing, improving, and enhancing the data and their metadata [1,23]. Furthermore, the curation process aims to ensure that the data and metadata meet a defined set of quality requirements, such as security rules, integrity constraints, or metadata availability expectations. Without proper curation, data resources may deteriorate in terms of their quality and integrity over time. One of the major challenges for achieving efficient and continuous curation of metadata is to create a methodology to structure the curation process as well as to provide a set of tools to support the curation process.

6 Conclusions

The way that individuals and organizations have produced, shared, and consumed data has changed with the advent of new technologies. As a consequence, data have become a tradable and valuable good. There are now Data Ecosystems, in which communities of actors interact with each other to exchange, produce, and consume data. The Data Ecosystem is an area that has been gaining popularity in the last five years. This article takes the form of a systematic mapping of the literature on the Data Ecosystem field. We found and analyzed 29 relevant studies from a gross total of 250 extracted from a list of online library databases.

In conclusion, we identify the field of Data Ecosystems as a new field of growing importance. However, there are some gaps in the literature that may hinder the development of Data Ecosystems. This study reveals three important areas that should be considered in order to add greater detail on how to develop and evolve the Data Ecosystem field: theory, models, and engineering. So far, the role of Data Ecosystem theory is not well defined. Without a good common understanding of this theory, it is difficult to develop an effective Data Ecosystem and its related technologies. As to Data Ecosystem models, the current proposed models cover only a small fragment of how a Data Ecosystem works. To the best of our knowledge, there is a lack of common conceptual models and health assessment models. Finally, there is also a lack of engineering methods to provide a common structure in the form of well-defined rules, procedures, protocols, and processes to develop, manage, and evolve Data Ecosystems.

References

1. Abbott D (2013) What is digital curation. http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation/. Accessed 20 Jan (2018)
2. Abu-Matar M (2016) Towards a software defined reference architecture for smart city ecosystems. In: 2016 IEEE international smart cities conference (ISC2). IEEE, pp 1–6
3. Alves C, Oliveira J, Jansen S (2017) Software ecosystems governance—a systematic literature review and research agenda. In: ICEIS 2017-proceedings of the 19th international conference on enterprise information systems, vol 3. pp 26–29
4. Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. Int J Soc Res Methodol 8:19–32
5. Atkinson C, Kuhne T (2003) Model-driven development: a metamodeling foundation. IEEE Softw 20(5):36–41
6. Attard J, Orlandi F, Auer S (2016). Data value networks: enabling a new data ecosystem. In: 2016 IEEE/WIC/ACM international conference on web intelligence (WI). IEEE, pp 453–456
7. Barbosa L, Pham K, Silva C, Vieira MR, Freire J (2014) Structured open urban data: understanding the landscape. Big Data 2(3):144–154
8. Barnaghi P, Sheth A, Henson C (2013) From data to actionable knowledge: big data challenges in the web of things. IEEE Intell Syst 28(6):6–11
9. Bourne PE, Lorsch JR, Green ED (2015) Perspective: sustaining the big-data ecosystem. Nature 527(7576):S16–S17
10. Campbell PR, Ahmed F (2010) A three-dimensional view of software ecosystems. In: Proceedings of the fourth European conference on software architecture: companion volume. ACM, pp 81–84
11. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171–209
12. Christensen HB, Hansen KM, Kyng M, Manikas K (2014) Analysis and design of software ecosystem architectures-towards the 4S telemedicine ecosystem. Inf Softw Technol 56(11):1476–1492
13. Chun SA, Shulman S, Sandoval R, Hovy E (2010) Government 2.0: making connections between citizens, data and government. Inf Polity 15(1):1
14. Chyi Lee C, Yang J (2000) Knowledge value chain. J Manag Dev 19(9):783–794
15. Clegg CW (2000) Sociotechnical principles for system design. Appl Ergon 31(5):463–477
16. Daniel EM, Wilson HN (2003) The role of dynamic capabilities in e-business transformation. Eur J Inf Syst 12(4):282–296

17. da Silva AR (2015) Model-driven engineering: a survey supported by the unified conceptual model. Comput Lang Syst Struct 43:139–155
18. Davies T (2011) Open data: infrastructures and ecosystems. In: Proceedings of web science conference, pp 1–6
19. Dawes SS, Vidiasova L, Parkhimovich O (2016) Planning and designing open government data programs: an ecosystem approach. Gov Inf Q 33(1):15–27
20. Ding L, Lebo T, Erickson JS, Difranzo D, Williams GT, Li X, Michaelis J, Graves A, Zheng JG, Shangguan Z et al (2011) TWC LOGD: a portal for linked open government data ecosystems. Web Semant Sci Serv Agents World Wide Web 9(3):325–333
21. Fischer G, Herrmann T (2011) Socio-technical systems. In: Knowledge and technological development effects on organizational and social structures. IGI Global, pp 1–36. https://doi.org/10.4018/978-1-4666-2151-0.ch001
22. Fisher DM (2004) The business process maturity model: a practical approach for identifying opportunities for optimization. BPTrends, pp 1–7
23. Freitas A, Curry E (2016) Big data curation. In: Cavanillas JM, Curry E, Wahlste W (eds) New horizons for a data-driven economy. Springer, pp 87–118
24. Gama K, Lóscio BF (2014) Towards ecosystems based on open data as a service. In: ICEIS (2). pp 659–664
25. Group OGW (2007) Eight principles of open government data, *Open Government Working Group*. https://public.resource.org/8_principles.html. Accessed 20 Jan 2018
26. Ha S, Lee S, Lee K (2014) Standardization requirements analysis on big data in public sector based on potential business models. Int J Softw Eng Its Appl 8(11):165–172
27. Hanssen GK (2012) A longitudinal case study of an emerging software ecosystem: implications for practice and theory. J Syst Softw 85(7):1455–1466
28. Hanssen GK, Dybå T (2012) Theoretical foundations of software ecosystems. In: IWSECO@ ICSOB. Citeseer, pp 6–17
29. Harrison TM, Pardo TA, Cook M (2012) Creating open government ecosystems: a research and development agenda. Future Internet 4(4):900
30. Heimstädt M, Saunderson F, Heath T (2014) Conceptualizing open data ecosystems: a timeline analysis of open data development in the UK. In: CeDEM14: conference for E-democracy an open government. MV-Verlag, p 245
31. Helbig N, Cresswell AM, Burke GB, Luna-Reyes L (2012) The dynamics of opening government data. Center for Technology in Government.[Online]. http://www.ctg.albany.edu/publications/reports/opendata. Accessed 20 Jan 2018
32. Iansiti M, Levien R (2004) The keystone advantage: what the new dynamics of business ecosystems mean for strategy, innovation, and sustainability. Harvard Business Press, Brighton
33. Immonen A, Palviainen M, Ovaska E (2014) Requirements of an open data based business ecosystem. IEEE Access 2:88–103
34. Jalali S, Wohlin C (2012) Systematic literature studies: database searches vs. backward snowballing. In: Proceedings of the ACM-IEEE international symposium on empirical software engineering and measurement. ACM, pp 29–38
35. Jansen S, Brinkkemper S, Finkelstein A (2009) Business network management as a survival strategy: a tale of two software ecosystems. In: Proceedings of the 1st international workshop on software ecosystems. p 34
36. Jansen S, Cusumano MA, Brinkkemper S (2013) Software ecosystems: analyzing and managing business networks in the software industry. Edward Elgar Publishing, Cheltenham
37. Janssen M, Charalabidis Y, Zuiderwijk A (2012) Benefits, adoption barriers and myths of open data and open government. Inf Syst Manag 29(4):258–268
38. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Tech. rep., School of Computer Science and Mathematics, Keele University
39. Köster V, Suárez G (2016) Open data for development: experience of uruguay. In: Proceedings of the 9th international conference on theory and practice of electronic governance. ACM, pp 207–210
40. Koznov D, Andreeva O, Nikula U, Maglyas A, Muromtsev D, Radchenko I (2016) A survey of open government data in Russian Federation. In: IC3K 2016—Proceedings of the 8th international joint conference on knowledge discovery, knowledge engineering and knowledge management, vol 3. pp 173–180
41. Lee D (2014) Building an open data ecosystem: an Irish experience. In: Proceedings of the 8th international conference on theory and practice of electronic governance. ACM, pp 351–360
42. Lettner D, Angerer F, Prähofer H, Grünbacher P (2014) A case study on software ecosystem characteristics in industrial automation software. In: Proceedings of the 2014 international conference on software and system process. ACM, pp 40–49

43. Lindman J, Kinnari T, Rossi M (2016) Business roles in the emerging open-data ecosystem. IEEE Softw 33(5):54–59
44. Lopez-Herrejon RE, Linsbauer L, Egyed A (2015) A systematic mapping study of search-based software engineering for software product lines. Inf Softw Technol 61:33–51
45. Lundell B, Forssten B, Gamalielsson J, Gustavsson H, Karlsson R, Lennerholt C, Lings B, Mattsson A, Olsson E (2009) Exploring health within OSS ecosystems. In: First international workshop on building sustainable open source communities (OSCOMM 2009), Skövde, Sweden, pp 1–5
46. Madhavan J, Jeffery SR, Cohen S, Dong XL, Ko D, Yu C, Halevy A (2007) Web-scale data integration: you can only afford to pay as you go. In: Conference on innovative data systems research
47. Magalhaes G, Roseira C, Manley L (2014) Business models for open government data. In: Proceedings of the 8th international conference on theory and practice of electronic governance. ACM, pp 365–370
48. Manikas K, Hansen KM (2013) Reviewing the health of software ecosystems—a conceptual framework proposal. In: Proceedings of the 5th international workshop on software ecosystems (IWSECO). pp 33–44
49. Manikas K, Hansen KM (2013) Software ecosystems—a systematic literature review. J Syst Softw 86(5):1294–1306
50. Mercado-Lara E, Gil-Garcia JR (2014) Open government and data intermediaries: the case of AidData. In: Proceedings of the 15th annual international conference on digital government research. ACM, pp 335–336
51. Moiso C, Minerva R (2012) 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 (ICIN). IEEE, pp 202–209
52. Moore JF (1999) Creating value in the network economy. pp 121–141. http://dl.acm.org/citation.cfm?id=303444.303452. Accessed 20 Jan 2018
53. Munro R (2009) Actor-network theory. The SAGE handbook of power. Sage Publications Ltd, London, pp 125–39
54. Nachira F, Dini P, Nicolai A (2007) A network of digital business ecosystems for Europe: roots, processes and perspectives. European Commission, Bruxelles, Introductory Paper
55. Oliveira MIS, de Oliveira HR, Oliveira LA, Lóscio BF (2016) Open government data portals analysis: the Brazilian case. In: Proceedings of the 17th international digital government research conference on digital government research. ACM, pp 415–424
56. Oliveira Lairson Alencar OMIS, Santos WCdR, Lóscio BF (2018) Data on the web management system: a reference model. In: Proceedings of the 19th international digital government research conference on digital government research. ACM
57. Ordanini A, Pol A (2001) Infomediation and competitive advantage in B2B digital marketplaces. Eur Manag J 19(3):276–285
58. Organization T (2014) Sustainability. www.thwink.org/sustain/glossary/Sustainability.htm. Accessed 20 Jan 2018
59. Palvia P, Leary D, Mao E, Midha V, Pinjani P, Salam A (2004) Research methodologies in MIS: an update. Commun Assoc Inf Syst 14(1):24
60. Petersen K, Feldt R, Mujtaba S, Mattsson M (2008) Systematic mapping studies in software engineering. In: EASE, vol 8. pp 68–77
61. Pfeffer J, Salancik GR (2003) The external control of organizations: a resource dependence perspective. Stanford University Press, Stanford
62. Poikola A, Kola P, Hintikka K (2011) Public data, an introduction to opening information resources. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/78201/Public_data_-_an_introduction_to_opening_information_resources.pdf?sequence=1. Accessed 20 Jan 2018
63. Pollock R (2011) Building the (open) data ecosystem. In: Open knowledge foundation Blog 31. https://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/. Accessed 20 Jan 2018
64. Sandelowski M (2010) What's in a name? qualitative description revisited. Res Nurs Health 33(1):77–84
65. Shaw M (2011) Writing good software engineering research papers. In: 25th international conference on software engineering, 2003. Proceedings. IEEE, pp 726–736
66. Shin DH (2016) Demystifying big data: anatomy of big data developmental process. Telecommun Policy 40(9):837–854
67. Shin DH, Choi MJ (2015) Ecological views of big data: perspectives and issues. Telemat Inform 32(2):311–320
68. Silver B, Richard B (2009) BPMN method and style, vol 2. Cody-Cassidy Press Aptos, Aptos
69. Sjoberg DI, Dyba T, Jorgensen M (2007) The future of empirical methods in software engineering research. In: Future of software engineering, 2007. FOSE'07. IEEE, pp 358–378

70. Smith G, Ofe HA, Sandberg J (2016) Digital service innovation from open data: exploring the value proposition of an open data marketplace. In: 2016 49th Hawaii international conference on system sciences (HICSS). IEEE, pp 1277–1286
71. Sommerville I (2010) Software engineering. Pearson, London
72. Sooklal R, Papadopoulos T, Ojiako U (2011) Information systems development: a normalisation process theory perspective. Ind Manag Data Syst 111(8):1270–1286
73. Stol KJ, Fitzgerald B (2015) Theory-oriented software engineering. Sci Comput Program 101:79–98
74. Teece DJ (2010) Business models, business strategy and innovation. Long Range Plan 43(2–3):172–194
75. Ubaldi B (2013) Open government data: towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance. https://doi.org/10.1787/19934351
76. Van Schalkwyk F, Willmers M, McNaughton M (2016) Viscous open data: the roles of intermediaries in an open data ecosystem. Inf Technol Dev 22(sup1):68–83
77. Welle Donker F, van Loenen B (2017) How to assess the success of the open data ecosystem? Int J Digit Earth 10(3):284–306
78. Wikipedia (2001) Ecosystem https://en.wikipedia.org/wiki/Ecosystem. Accessed 20 Jan 2018
79. Yoo Y, Henfridsson O, Lyytinen K (2010) Research commentary-the new organizing logic of digital innovation: an agenda for information systems research. Inf Syst Res 21(4):724–735
80. Zeleti FA, Ojo A (2014) Capability matrix for open data. In: Working conference on virtual enterprises. Springer, pp 498–509
81. Zeleti FA, Ojo A (2016) Open data value capability architecture. Inf Syst Front 19(2):1–24
82. Zeleti FA, Ojo A (2016) Critical factors for dynamic capabilities in open government data enabled organizations. In: Proceedings of the 17th international digital government research conference on digital government research - dg.o '16. ACM Press. https://doi.org/10.1145/2912160.2912164
83. Zott C, Amit R (2010) Business model design: an activity system perspective. Long Range Plan 43(2–3):216–226
84. Zubcoff JJ, Vaquer L, Mazón JN, MaciÁ F, Garrigós I, Fuster A, Carcel JV (2016) The university as an open data ecosystem. Int J Des Nat Ecodyn 11(3):250–257
85. Zuiderwijk A, Janssen M (2014) Barriers and development directions for the publication and usage of open data: a socio-technical view. In: Gascó-Hernández M (ed) Open government. Springer, pp 115–135
86. Zuiderwijk A, Janssen M, Choenni S, Meijer R, Alibaks RS, Sheikh_Alibaks R (2012) Socio-technical impediments of open data. Electron J e-Gov 10(2):156–172
87. Zuiderwijk A, Janssen M, Davis C (2014) Innovation with open data: essential elements of open data ecosystems. Inf Polity 19(1, 2):17–33
88. Zuiderwijk A, Janssen M, van de Kaa G, Poulis K (2016) The wicked problem of commercial value creation in open data ecosystems: policy guidelines for governments. Inf Polity 21(3):223–236
89. Zuiderwijk A, Janssen M, Poulis K, van de Kaa G (2015) Open data for competitive advantage: insights from open data use by companies. In: Proceedings of the 16th annual international conference on digital government research. ACM, pp 79–88

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Marcelo Iury S. Oliveira received his Bachelor's degree in computing science from the State University of Ceará, Brazil, in 2005. He also received his M.Sc. degree in computing science from the Federal University of Campina Grande, Brazil, in 2007. He is currently a Ph.D. candidate in computing science from Federal University of Pernambuco, Brazil. Moreover, he is currently a Professor in the Federal Rural University of Pernambuco. His main research areas are Distributed Systems, Software Engineering, Databases, and Data Ecosystems.

Glória de Fátima Barros Lima received her Bachelor's degree in Computer Engineering from Federal University of Pernambuco, Brazil, in 2016. She is currently a master's candidate in Computer Science from Federal University of Pernambuco. Her main research interests are focused on Software Testing, Agile Methodologies, Big Data, Databases, and Data Ecosystems.

Bernadette Farias Lóscio received her M.Sc. degree in Computer Science from Federal University of Ceará and her Ph.D. in Computer Science from Federal University of Pernambuco in 1998 and 2003, respectively. Currently, she is an Associate Professor at the Federal University of Pernambuco. Her main research interests are mainly focused on Data on the Web, Open Data, Data Ecosystems, Semantic Web, Data Integration, and Big Data.