IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 22, 2020, accepted January 3, 2021, date of publication January 18, 2021, date of current version January 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052025

Systematic Mapping of Open Data Studies: Classification and Trends From a Technological Perspective

ROBERT ENRÍQUEZ-REYES[1], SUSANA CADENA-VELA[1], ANDRÉS FUSTER-GUILLÓ[2], JOSE-NORBERTO MAZÓN[2], LUIS DANIEL IBÁÑEZ[3], AND ELENA SIMPERL[4]

[1]Facultad de Ciencias Administrativas/Facultad de Ingeniería y Ciencias Aplicadas, Universidad Central del Ecuador, Quito 170521, Ecuador
[2]Instituto Universitario de Investigación Informática, Universidad de Alicante, 03690 Alicante, Spain
[3]Department of Electronic and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.
[4]Department of Informatics, King's College London, London WC2B 4BG, U.K.

Corresponding author: Jose-Norberto Mazón (jnmazon@ua.es)

This work was supported by the Spanish Government (MINECO) under Project TIN2016-78103-C2-2-R and Project TIN2017-89069-R, and in part by the European Union Horizon 2020 Program through the ODINE Project under Grant 644683 and the Data Pitch Project under Grant 732506. The work of Jose-Norberto Mazón was supported by the "José Castillejo" research program from the Spanish Government (Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+i, Subprograma de Movilidad, del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016) under Grant JC2015-00284.

ABSTRACT The objective of this paper is to classify and analyse all research on open data performed in the scientific community from a technological viewpoint, providing a detailed exploration based on six key facets: publication venue, impact, subject, domain, life-cycle phases and type of research. This paper therefore provides a consolidated overview of the open data arena that allows readers to identify well-established topics, trends, and open research issues. Additionally, we provide an extensive qualitative discussion of the most interesting findings to pave the way for future research. Our first identification phase resulted in 893 relevant peer-reviewed articles, published between 2006 and 2019 in a wide variety of venues. Analysis of the results shows that open data research grew slowly from 2006 but increased significantly as from 2009. In 2019, research interest in open data from a technological perspective overall decreased. This fact could indicate that research is beginning to stabilise, i.e., the open data research hype is over, and the research field is reaching maturity. Main findings are (i) increasing effort in researching on Semantic Web technologies as a mechanism to publish and reuse linked open data, (ii) software systems are proposed to solve open data technical problems; and (iii) considering technological aspects of legislation and standardization is needed to widely introduce open data in society. Finally, we provide complementary insights regarding open data innovation projects, with special emphasis on publication (e.g., open data portals) and consumption (e.g., open data as business enabler) of open data.

INDEX TERMS Open data, systematic mapping, innovation.

I. INTRODUCTION

The concept of open data (i.e., data freely used, modified, and shared by anyone for any purpose[1]) emerged in the early 2000s with some relevant milestones as the 2003 European Public-Sector Information (PSI) [1], or the 2009 United States Government decision to implement the Open Government concept [2], which had a worldwide impact. Open data has also been driven by international organizations such as the World Bank (2012) [3]. In 2013, the G8 group formed by world leaders signed the open data Charter [4], aiming at fostering broader global adoption of open data. Due to these global initiatives, the open data term gained momentum and the body of research on open data began to emerge as a multidisciplinary area encompassing a wide range issues, from social to technical. Research on open data is intrinsically multidisciplinary mainly due to a couple of facts:

- Open data is published in the Web, as an information space where social and technical aspects come together [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.

[1] https://okfn.org/opendata/

- Open data is published by following some data quality criteria (such as the 5-star schema created by Berners-Lee), but also some legal guidelines (such as standard licenses), which both allow open data to be effectively reused and have a positive impact on society [6].

Even though, there are some papers that survey research on social aspects of open data initiatives, e.g. to classify open government data initiatives [1], there is a lack of works that review and classify open data research from a technological viewpoint (i.e., considering papers mainly coming from repositories on computer science and engineering).

The main goal of this paper is thus to provide a comprehensive review and classification of the open data field from a technological perspective, considering issues such as: domain, topic, impact, research type, venues, etc. To achieve this, we propose using a systematic mapping study. This method arose in the medical field [7] and has been used extensively in software engineering [8]. Systematic mapping studies provide a repeatable method aimed at performing a comprehensive overview of a research field, providing a useful reference for further researchers. According to [9], the open data systematic mapping presented in this paper can be a useful resource for:

- Beginning technological researchers in the open data field. Classification of the conducted research on open data is described in our systematic mapping, thus giving valuable insights for starting research.
- Experienced researchers, who can use this document as a qualitative reference work for subsequent studies. Our mapping study provides understanding of the existing literature on specific topics in the open data arena and allows to identify the need for additional research in specific areas.
- Industrial actors, such as data publishers or data reusers, who need a thorough introduction of the open data field from a technological perspective. Our mapping study allows these actors to get an overview of the state-of-the-art and to identify trends and clusters of open data research studies that are suitable and applicable for them, aiding communication and knowledge transfer between academia and industry.

Furthermore, opening data favours the development of innovation, contributing to the improvement of efficiency [5]. Therefore, a review of open data innovation projects (from insights of the systematic mapping) is also provided.

In this paper, we begin by explaining in Section II the research method and process we followed: definition of research scope, identification of papers with required conditions, specification of the classification scheme, and classification of publications. In Section III, we analyse the extracted data and visualize the results using stacked bar, pie, and bubble charts. Section IV provides a discussion of the results revealing interesting insights into the open data research field. We provide an in-depth qualitative analysis of the most remarkable results of the systematic mapping study, summarizing the areas of study and most relevant worldwide applications. Section V surveys some research insights regarding open data innovation projects, and finally, Section VI provides conclusions.

II. RESEARCH METHOD

There are different methods for reviewing research [10], ranging from narrative literature review to systematic approaches guided by a replicable process. Narrative reviews aim to appraise previous research without describing a formal process for identifying, selecting, and evaluating relevant publications (which may produce biased results). Regarding systematic methods, there are several to be used depending on the pursued goals:

- Systematic reviews aim to select and appraise all available research depending on some research questions. However, the focus is not on classifying research.
- Rapid reviews speed up the systematic review process for updating previous reviews or for considering novel emerging topics.
- Scoping reviews aim to select and classify existing research literature in terms of topics (with no research questions that guide the review process).
- Systematic mappings are focused on performing a visual synthesis of the research publications and classify them based on some facets provided by research questions. Systematic mappings are best designed for a scenario where there are an abundance and a diversity of research in order to identify gaps in a research area. Therefore, it is the ideal method for classifying research on open data.

A systematic mapping study aims at finding and classifying primary studies in a specific topic area by following a well-defined and repeatable method [9]. It consists of several steps, namely: (i) obtaining a classification scheme, (ii) gathering relevant studies, (iii) performing the classification, and (iii) analysing the results [7]. The analysis focuses on answering specific research questions, usually related to the identification and coverage of the field and its subfields, and on the evolution over time, as well as additional discussion on challenges and trends regarding the specific topic, as stated by [11].

For our systematic mapping study, we deployed the process introduced by [11] which was inspired by the adaptation of systematic mapping studies in the medical field and their application to software engineering proposed by [7]. Our process thus consisted in: defining the scope of the research, defining the search process, the classification scheme, the mapping of publications according to the classification scheme, and data analysis. A detailed overview of this process is shown in Figure 1. The methodology of the systematic mapping study is as follows:

- First step (research scope): definition of the research questions that will guide the research.

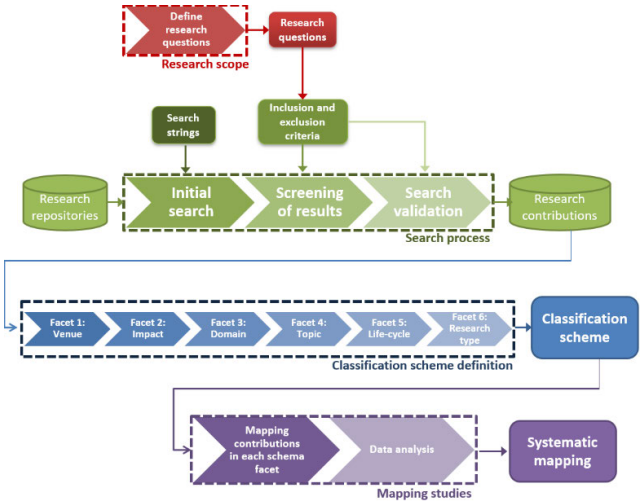

FIGURE 1. Research process for the conducted systematic mapping.

- Second step: the search process. After selecting the search strings and the inclusion and exclusion criteria, a search (including screening of the results and the validation of the search) was conducted to obtain results from research repositories.
- Third step: definition of the classification scheme aiming at defining the facet of each research category (specifically, six facets were defined for our systematic mapping study).
- Last step: mapping of studies. This step consists of obtaining a classification of the studies, manually performed by the authors, according to the classification schema as well as performing a detailed analysis of the data.

A. RESEARCH SCOPE

The overall goal of this systematic mapping study is to provide a consolidated overview of research in the field of open data from a technological perspective, through its publication venues, impact domains, phases of data publication, impact level, and their evolution over time. The development process answered the following research questions (RQs).

- RQ1: What are the publication venues in which open data research has been published?
- RQ2: What impact does the research on open data have?
- RQ3: Which domains received most coverage in open data research, to what extent, and how is coverage evolving?
- RQ4: What are the topics being addressed in open data research, to what extent are they covered, and how is coverage evolving?
- RQ5: Which phases of the data life-cycle have been considered in the open data research?
- RQ6: What types of open data research have been reported, to what extent, and how is the evolution progressing?

TABLE 1. Scientific repositories.

Name	URL
Springer RD research & development	http://rd.springer.com/
IEEE Xplore	http://ieeexplore.ieee.org/.
ACM Digital Library	http://dl.acm.org
IOS Press	http://content.iospress.com
Science Direct	http://www.sciencedirect.com/

B. SEARCH PROCESS

The search process aims at identifying primary studies that are relevant within the research scope in a well-defined, repeatable way. A three-step search process was carried out to identify such primary studies.

1) STEP 1: INITIAL SEARCH

To obtain the initial set of potentially relevant primary studies (i.e. "publications"), an extensive search was performed using search engines from the most important scientific repositories of publishers within the broad field of information technology (in accordance with the technological perspective of our study); therefore, we selected the search engines included in Table 1.

All searches were performed between 1 September 2016 and 20 December 2019. Since our goal was to study the publications on open data and how this term has been specifically used in the research area, publication titles had to include the specific search string: "open data". Publications between 2006 and 2019 were included.

We only took into account publications in the English language. General search engines for scientific publications such as Google Scholar were not considered for two reasons: (i) they index publications from the aforementioned repositories; and (ii) they list a lot of grey literature (i.e., research produced outside traditional academic publishing and distribution channels such as reports, working papers, government documents, white papers and so on).

The search engines produced 893 search results: these were classified, quantified, and presented in the results analysis, based on the different facets derived from the research questions.

2) STEP 2: SCREENING OF RESULTS

The initial set of publications contained irrelevant results that had to be discarded. In our screening process, we began by eliminating duplicates. Next, we examined the relevance of each publication with respect to our research objectives. To do so, we defined the inclusion and exclusion criteria. Inclusion and exclusion criteria were established in both form and content:

1) Form

- Inclusion criteria: all peer-reviewed publications in English for which the full text was available. This included all short and full research papers

published in peer-reviewed journals, conferences, symposia, or workshops.
- Exclusion criteria: sources that did not go through a peer-review process or did not constitute pure research contributions, for example: books, Ph.D. and master's theses, patent descriptions, standards, and recommendations, book or thesis summaries, technical reports, white papers, invited talks, demo papers, tutorial papers, poster publications, posters, editorials, prefaces, articles or columns in non-peer-reviewed journals, newsletters, encyclopaedia entries, summaries or blog posts. We further excluded sources for which the full text was not published, such as abstracts, extended abstracts, and presentations (slideshows).

2) Content

- Inclusion criteria: the abstract or introduction explicitly mentions open data, as well as it can be deduced that the research was explicitly performed for solving a problem regarding publishing or reusing open data.
- Exclusion criteria: documents that did not focus on the publication or reuse of open data, for example those related to the publication of institutional data for internal users or those that used open data in some experiments, i.e. the content did not identify a research problem related to the reuse or publication of open data, but only incidentally mentioned open data.

As studying the evolution over time is one of the objectives, we did not restrict our search based on publication year. As a starting point we consider 2006, the year "open data" term arises in papers coming from repositories we considered; as endpoint we took 2019 (included), the last full year prior to the development of this study. As such, we cover 14 years of research on open data from a technological perspective.

The screening process was performed rigorously, applying the inclusion and exclusion criteria described above.

3) STEP 3: SEARCH VALIDATION

Three steps were undertaken to validate the aforementioned search process in order to ensure our screening process produces a complete set (i.e., identifying missing publications). Inclusion and exclusion criteria of the screening process are reapplied.

1) We took the transitive closure of the relevant publications using their references.
2) For each research venue where more than one relevant publication was found (see research question 1 in subsection II.A), we checked all issues and proceedings (from 2006 until 2019).
3) We repeated all our previous searches and checked if those publications are in the scientific repositories.

Step 1 was performed iteratively, but no new results complying with the inclusion criteria were found. Also, step 2 and step 3 did not result in any new publications, indicating that our search process can be considered complete. This seems reasonable as our search term was rather generic.

C. CLASSIFICATION SCHEME DEFINITION

After obtaining the final set of publications, we devised a classification scheme corresponding to the research scope and questions set out in Section II.A. Based on these research questions, we considered six facets for classification: venue, impact, domain, topic, data lifecycle phase, and research type.

1) FACET 1. VENUE

Together with each publication coming from the scientific repositories after performing the screening process, we found information about the conference or scientific journal in which it is published. We also quantified the number of publications by conference or scientific journal.

2) FACET 2. IMPACT

Description: it describes the scope of the research results.

Classification scheme. It is divided into:

- Local impact: which describes a small and focused area the open data relates to, for example: a city, a museum, a university, a hospital.
- Regional impact: related to a larger area such as a province or a state.
- National impact: related to a specific country.
- International impact: related to an international area.

3) FACET 3. DOMAIN

Description: domains receiving coverage in open data research.

Derivation method: domains were borrowed from the 14 data categories suggested by the G8 Open Data Charter [12].

Classification scheme. It consists of the following domains:

- Agriculture: food security, farmers and final consumers, sustainable agricultural development, nutrition
- Biology: articles that deal with the study of life and living organisms.
- Chemical: articles that deal with matter, material composition and reaction.
- Culture: articles that deal with social behaviour and norms found in human societies.
- Data Journalism: approaches for storytelling and journalism, journalistic activities.
- Economy: financial movements, income, expenses and budget.
- Education: educational applications, school performance, digital skills.
- Environment: environment, climate change, natural resources, environmental information, forest landscapes, meteorology/climate, agriculture, forestry, fishing and hunting, pollution levels, energy consumption.

- Geospatial: topography, postal codes, national maps, local maps, food security, lakes, geography zones.
- Health: papers related to the study of the state of complete physical, mental and social wellbeing and not simply the absence of disease.
- Humanitarian: humanitarian assistance, disaster management, relief and reconstruction activities, disasters, epidemics.
- Infomediaries: workers, companies and investors who work in identifying and leveraging the market value of consuming information (see [13]).
- Innovation: original and more effective use of open data for making a meaningful impact in market (e.g., to boost economies) or society (e.g., to improve daily life of people).
- Science: scientific research and discoveries, innovative methods to open scientific data and create new tools to manipulate that data, financing of scientific projects, collaboration among stakeholder groups.
- Transport: public transport timetables, access points broadband penetration.
- Energy: energy development-oriented applications
- Tourism: open data applications applied to tourism development.

4) FACET 4. TOPIC

Description: topics being addressed in open data research.

Derivation method: topics were taken from the ''Call for Papers'' of two of the most important scientific conferences on data and the Semantic Web, namely, the International Conference on Very Large Data Bases (VLDB) and the International Semantic Web Conference (ISWC), respectively. Regarding these two conferences, a quantification was made and the most frequently recurring topics in the calls for papers of both conferences over the last three years (2017, 2018 and 2019) were established.

Classification scheme. Topics were classified according to:

- Entrepreneurship: entrepreneurial usage of open data. Use and impact of open data in specific countries or specific sectors in order to leverage business and the economy.
- Government: the making of, implementation and institutionalization of open data policy, capacity-building for wider availability and use of open data, conceptualizing open data ecosystems and intermediaries; linkages between transparency, freedom of information and open data communities; measurement of open data policy and practices, including methods for assessing the impact of open data; situating open data in the global governance and development context.
- Information retrieval: databases, information retrieval, information extraction, natural language processing for searching and querying databases, fuzzy, probabilistic, and approximate databases, information retrieval, text in databases.
- Infrastructure: cloud environments for data, access methods, concurrency control, recovery, transactions, indexing and search, in-memory data management, hardware accelerators, query processing and optimization, storage management, tuning, benchmarking, performance measurement, database administration and manageability, database-as-a-service.
- Intelligent systems: artificial intelligence, social networks, recommendation systems, business intelligence and data mining.
- Internet of things (IoT): data streams and the internet of things, crowdsourcing, embedded and mobile databases, real-time databases, sensors and IoT, stream databases.
- Quality: cleaning, quality assurance, and provenance of semantic web data, services, and processes, data cleaning, information filtering and dissemination, information integration, metadata management, data discovery, web data management, heterogeneous and federated database systems, database usability.
- Security: trust, privacy and security in data management, critical challenges for data: exclusion and abuse.
- Semantic web: knowledge graph creation, reasoning, usage, knowledge representation and reasoning on the web, scalable management of semantics and data on the web, including linked data, semantic web data analysis, languages, tools, and methodologies for representing and managing semantics and data on the web, architectures and algorithms for extreme volumes, heterogeneity, dynamicity, and decentralization of semantic web data, ontology-based data access and integration/exchange on the web, ontology engineering and ontology patterns for the web, ontology modularity, mapping, merging, and alignment for the web, supporting multilingualism in the semantic web, user interfaces and interaction with semantics and data on the web, information visualization and exploratory analysis methods for semantic web data, personalized access to semantic web data and applications, social semantics methods and applications.
- Software engineering: development of mobile platforms, distributed database systems, cloud data management, development of NoSQL databases, scalable analytics, distributed transactions, consistency, p2p and networked data management, software development and content delivery networks.
- Visualization: data models and query languages for visualization, schema management and design, user interfaces and visualization.

5) FACET 5. OPEN DATA LIFE-CYCLE

Description: the open data life-cycle describes the process of providing data as open data, i.e., preparing the data to be published, using the published data and curating the published data. Therefore, it is mainly concerned with three issues: preprocessing, exploitation and maintenance.

Derivation method: the following classification referred to the works by [1], [4], [12] which define an open government data life-cycle.

Classification scheme. The classification scheme consists of the following phases:

- Data creation: it refers to the generation of data as well as the collection of data for the specific purpose of publishing it.
- Data selection: this is the process involving selecting the data to be published. This requires removing any private or personal data, as well as identifying under which conditions this data will be published, potentially through the specification of open (government) data policies.
- Data harmonization: this step involves preparing the data to be published in order to conform to publishing standards.
- Data publishing: this is the specific act of opening up the data by publishing it on open data portals.
- Data interlinking: this is the final step in the 5-Star Scheme for Open Data (aforementioned), i.e. obtaining Linked Open Data. This allows published data to have additional value, as the linking of data gives context to it.
- Data discovery: the publishing of data is not enough to enable its reuse. Data consumers must discover the existence of open data in order to be able to consume it.
- Data exploration: this step is the most trivial way of consuming data. Here, a user passively examines open data by visualizing or scrutinizing it.
- Data exploitation: this step is a more advanced way of consuming data. Data exploitation enables a user to proactively use, reuse or distribute the open data by performing analysis, creating mashups, or innovating based on the open data.
- Data curation: While not necessarily occurring at a fix stage, data curation is vital in ensuring the published data is reusable. This involves a number of processes, including updating stale data, data and metadata enrichment, data cleansing, etc.

6) FACET 6. TYPE OF RESEARCH

Description: the research type is the type of reported research.

Derivation method: the type of research is not specific to the particular topic of open data but is generally applicable. As in the case of other systematic maps carried out in software engineering, we used the classification scheme proposed by [14] and [11].

Classification scheme. This scheme includes:

- Solution proposal: describes a solution usually illustrated with an example, case study, running example, etc. The work is barely or not validated (see next bullet point); the proposal is only explained and its application described.
- Validation research: validation of research that is not deployed in practice, for example, by an experiment, performing kinds of tests, lab studies, etc. Usually it follows a solution proposal. It answers the question: is the proposed solution ''good''?
- Evaluation research: an evaluation of research, usually by observing how the solution works in practice or comparing it with other solutions, pointing out positive and negative points. It is more extensive than validation and often carried out within an industrial setting. It answers the question: is the proposed solution the ''right'' solution?
- Philosophical or conceptual proposals: these sketch a new way of looking at existing things, providing a vision or philosophical view on a subject matter.
- Opinion paper: it describes the opinion of the authors, usually taking a positive or negative stance. It may also present an overview of a field or a comparison of techniques from the author's viewpoint. It is generally not based on related work or a research methodology.
- Experience paper: it describes the experience of the authors, usually in practice, using a certain method, technology, etc. Authors are usually people working in the industry [11].

D. MAPPING STUDIES

Based on the classification scheme mentioned above, the authors manually classified each of the 893 relevant publications according to each facet. To do so the authors used a spreadsheet in which they noted for each paper its type concerning each of the facets, getting recounts of the number of papers that fit into each type for each facet. To ensure correctness and consistency of the classification process, authors were divided into two groups and each of the 893 publications was assigned to one group to be reviewed and classified within each facet. In the case of conflicting results, the other group was asked to perform a classification and the results were discussed until reaching an agreement. A total of 10% of the 893 publications required discussion. This classification process is complex and costly. From the best of our knowledge, there are software tools to optimize the step of citation screening [15], [16], but the idea of using classification methods to automate the process could open up an unexplored line of work [17], [18].

E. THREATS TO VALIDITY

There are several factors that can threaten the validity of systematic mapping findings. In the literature [19], [20], the main deficiencies of systematic maps have been identified as follows: (i) bias in the selection of publications, (ii) errors in categorizing publications into detailed categories, and (iii) weak additional contribution from one publication to another (the so called ''delta papers'', i.e., papers that provide only minor additions compared to work previously published by the authors).

To mitigate the risk of the first threat, we started by working on a selection of search engines to cover the specific research area in depth and width. We selected specific search engines to cover the biggest publishers in the field of research (i.e. ACM, IEEE, Springer, IOS PRESS, ScienceDirect). Secondly, we had to make sure that all publications on the selected theme were found. To do this, we introduced the search validation step (see section 2.2) designed to identify missing publications after reviewing initial set of publications. Regarding the second identified threat, we followed a formal review process as stated above in this section.

Finally, to mitigate the risk of ''delta papers'', we grouped all publications based on authors such that all publications where at least 50% of the authors are the same are clustered together. Then detailed research contributions are summarized and compared with all other publications within the same cluster for possible delta papers. Possible delta papers are discussed among all the authors to come to a final decision. We found 6 delta papers out of a total of 893 (less than 1%), over all categories in each facet. Therefore, we considered this threat to be negligible and delta papers are not excluded from our study.

III. DATA ANALYSIS

In this section, we present the data analysis of the 893 publications, based on the research questions defined in Section II.A. Data can be found in the Zenodo open-access repository (https://zenodo.org/record/4433117). Different types of charts were used for conveniently answering the research questions:

- Stacked bar charts were used to represent the results per year, and analyse each of the facets discussed in Section II.C. They allow: (i) visualizing the most frequent publications and the frequency of the different categories within a facet, (ii) visualizing the relative weight of each facet, and (iii) being aware of the evolution of each facet category over time.
- Bar charts aim at visualizing the distribution of publications per venues and year.
- Bubble charts are used to show the relationships between the different facets and, as such, represent the systematic map(s) of open data research. These charts are traditionally used for this purpose because they allow three-dimensional representation of data, where each bubble represents the publication frequency with respect to two specific facets.

A. FACET 1. PUBLICATION VENUES

Figure 2 and Figure 3 show the venues that published most research related to open data for both journals and conferences, respectively. In Figure 2, we can identify 10 journals with 4 or more publications on open data, out of the 893 articles. The Semantic Web journal and the Government Information Quarterly journal were the most important journals

TABLE 2. Conference acronym.

Conference	Acronym
International Conference on Web Intelligence, Mining and Semantics	WIMS
Iberian Conference on Information Systems and Technologies	CISTI
IEEE Computer Society Computers, Software, and Applications Conference	COMPSAC
International Conference on Electronic Governance and Open Society: Challenges in Eurasia	EGOSE
IEEE International Geoscience and Remote Sensing Symposium	IGARSS
Hawaii International Conference on System Sciences	HICSS
International Conference Companion on World Wide Web	WWW
International Workshop on Open Data	WOD
International Conference on Theory and Practice of Electronic Governance	ICEGOV
International Digital Government Research Conference on Digital Government Research	dg.o

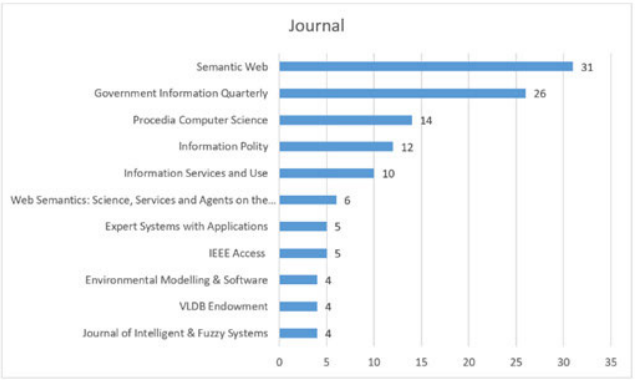

FIGURE 2. Journals ordered by number of publications.

for disseminating open data research from a technological perspective.

In Figure 3 we can identify proceedings of 10 conferences that most consider open data research. Worthy of note are the first two venues: DG.O (International Conference on Digital Government Research) and ICEGOV (Conference on Theory and Practice of Electronic Governance). Both conferences aim at considering open data research as multidisciplinary discipline, with emphasis on a technology point of view.

Figure 4 shows a pie chart representing the distribution of publications found in the source repositories defined according to facet 1 (venue). The scientific repository that contributed the most was the IEEE with 37%, followed by ACM with 33%. The others together represented 30% of publications. Remarkably, the first two scientific databases were those most related to IT, which confirms the relevance of a technological perspective of open data research.

Regarding the total number of publications contained in these most important scientific databases, we found out that IEEE contains more than 3 million papers in the period 2006-2019, while ACM contains just over 1,5 million papers.

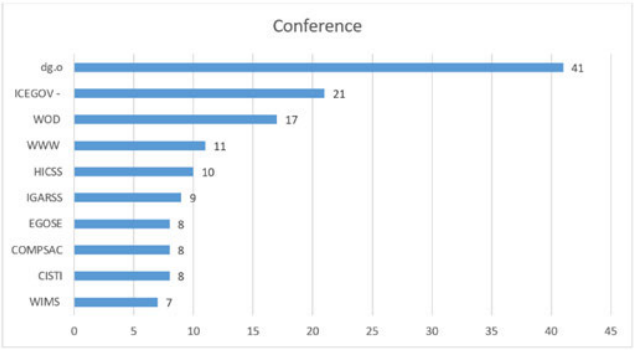

FIGURE 3. Conference proceedings ordered by number of publications.

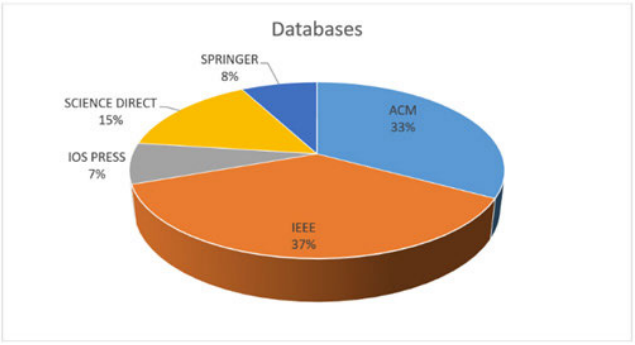

FIGURE 4. Distribution of publications in scientific databases.

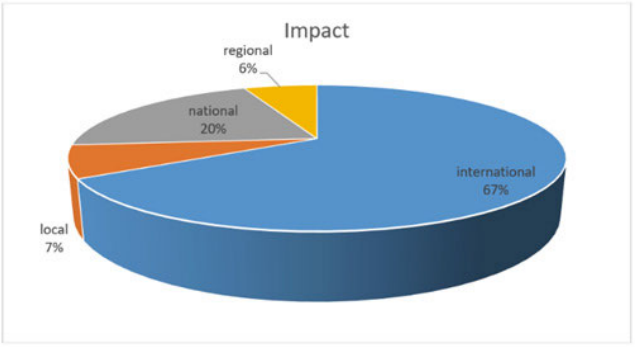

FIGURE 5. Percentages relating to publication impact.

Therefore, the number of papers on open data contained in the ACM scientific database is more significant than the number of papers on open data contained in the IEEE scientific database, as it represents a higher percentage of papers over the total. Furthermore, the IOS Press has the highest density of open data papers (i.e., number of open data papers from each database divided by the total number of papers from each database).

B. FACET 2. IMPACT

Figure 5 shows the percentages relating to publication impact according to facet 2 (impact): international publications were the most frequent with 67%.

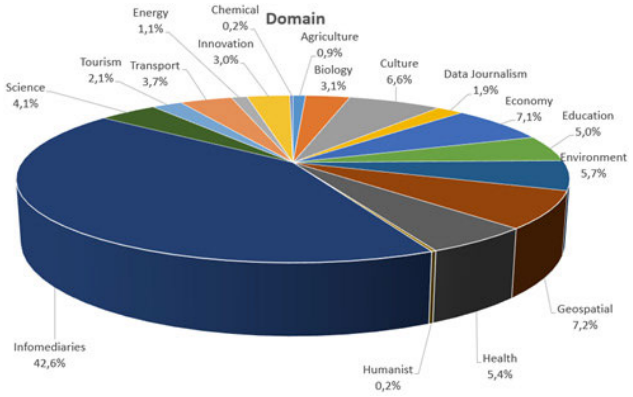

FIGURE 6. Percentages per domain.

C. FACET 3. DOMAIN

Figure 6 and Figure 7 show the distribution of publications according to the domain (facet 3) from 2006 and 2019. Aggregated values are shown in Figure 6, while Figure 7 provide results by year. ''Infomediaries'' were the most relevant domain with 42,6 % of publications. The other domains contributed less than 8% each, suggesting that publications mostly addressed problems from the open data consumers perspective.

Furthermore, Figure 7 shows that domains of ''transport'', ''economics'', ''education'', ''environment'', ''culture'', ''health'' and ''geospatial'' increased significantly in the same proportion over the past seven years; although there was no research in these domains from 2006 to 2009. It is worth noting that number of open data studies on ''innovation'' domain has increased in the last two years (2018 and 2019).

D. FACET 4. TOPIC

Figure 8 and Figure 9 illustrate the distribution of publications according to their topic (facet 4) and show how publications evolved from 2006 to 2019. The ''semantic web'' topic received most attention (22%), followed by ''software engineering'' (19%), and ''government'' (18%). The lack of research studies in open data quality (''quality'' topic) is surprising, mainly due to importance of having open data with enough quality to be properly reused.

Figure 9 shows that publications on ''Entrepreneurship'' are published from 2011 onwards, showing a positive trend until 2019. A significant number of publications on ''Software engineering'' are maintained throughout the period, although in 2015 there is a greater number. Also, the number of publications on ''Government'' is relevant during all period, but publications increased significantly in the last two years (2018 and 2019). Conversely, publications on ''Semantic Web'' were more relevant until 2018 but the number of publications on this topic decreases in 2019. The number of publications on ''IoT'' keeps stable since 2014, while ''Intelligent systems'' and ''Infrastructure'' topics become more relevant in the last two years (2018 and 2019).

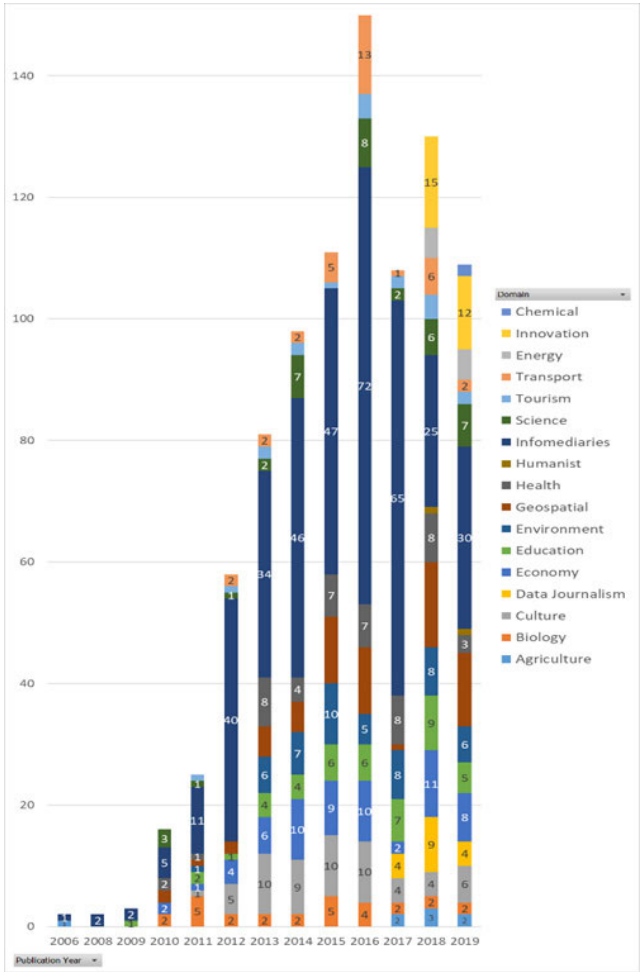

FIGURE 7. Distribution of publications according to the domain.

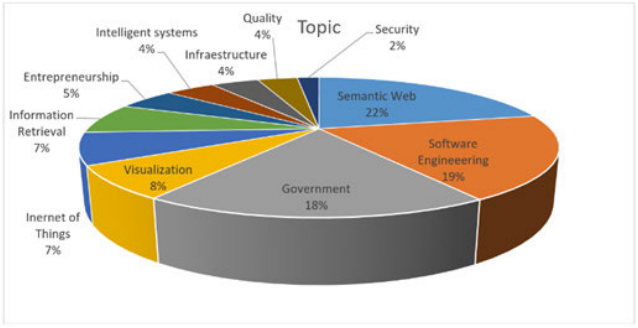

FIGURE 8. Percentages per topic.

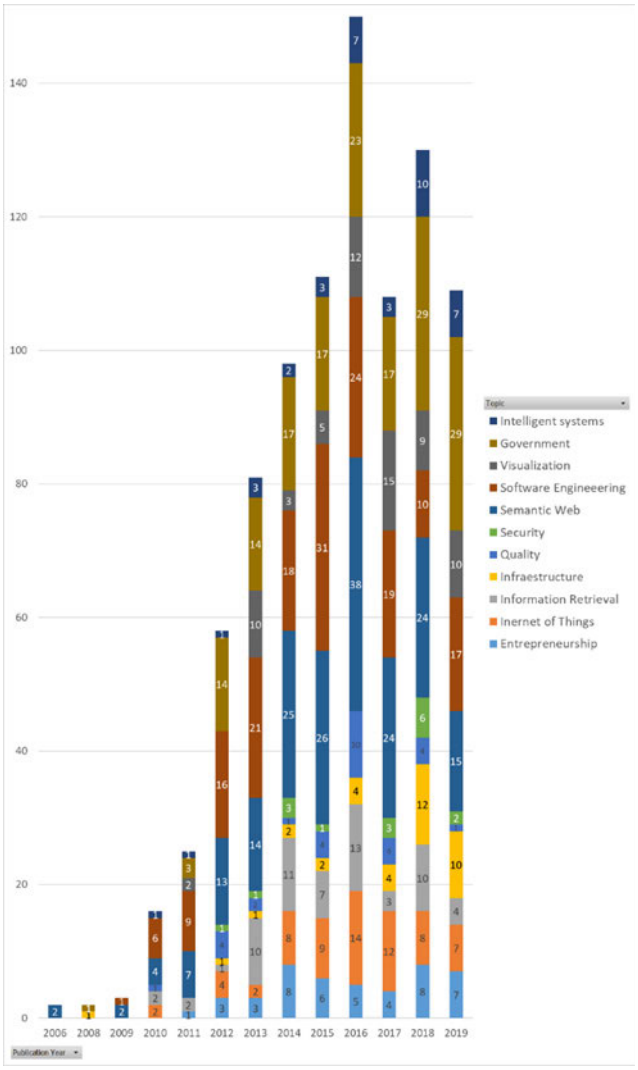

FIGURE 9. Distribution of publications per topic from 2006 to 2019.

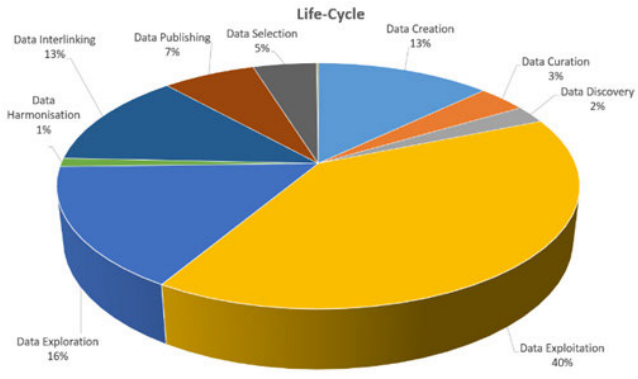

FIGURE 10. Pie chart representing the distribution of publications according to data life-cycle from 2006 to 2019.

E. FACET 5. DATA LIFE-CYCLE CLASSIFICATION

Figure 10 shows the distribution of the publications according to facet 5. Most studies focused on the ''data exploitation'' phase (40% of publications). The next phase was ''data exploration'' (16% of publications). This suggests that publications seeking to extract value from data and applicable results multiplied.

Figure 11 shows the distribution of publications along time, according to the phases defined in the life-cycle (facet 5). The ''exploitation'' phase grew remarkably until 2016. In 2009 it

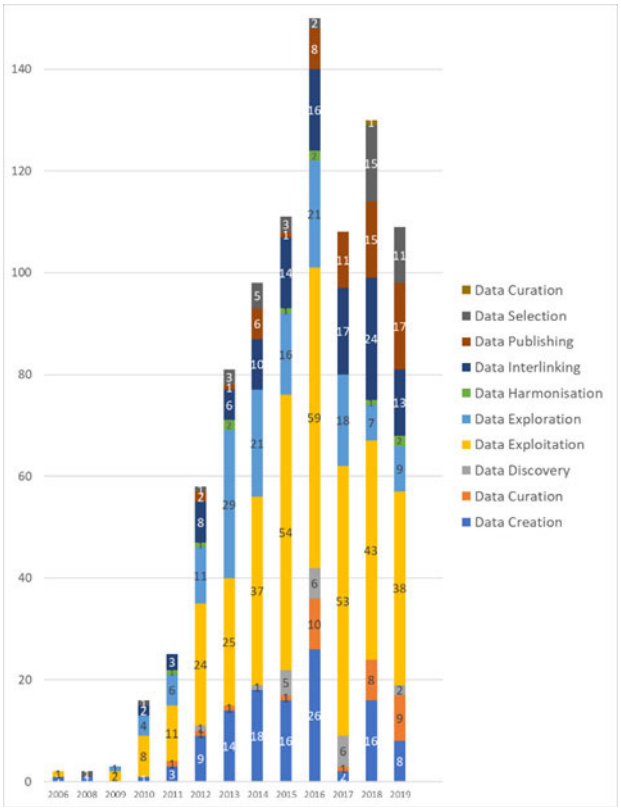

FIGURE 11. Distribution of publications according to data life-cycle from 2006 to 2019.

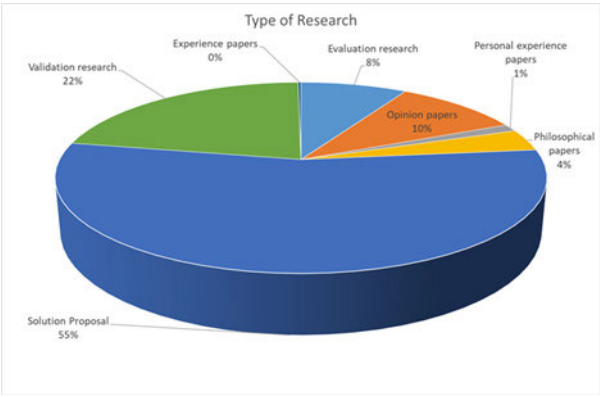

FIGURE 12. Percentages per type of research.

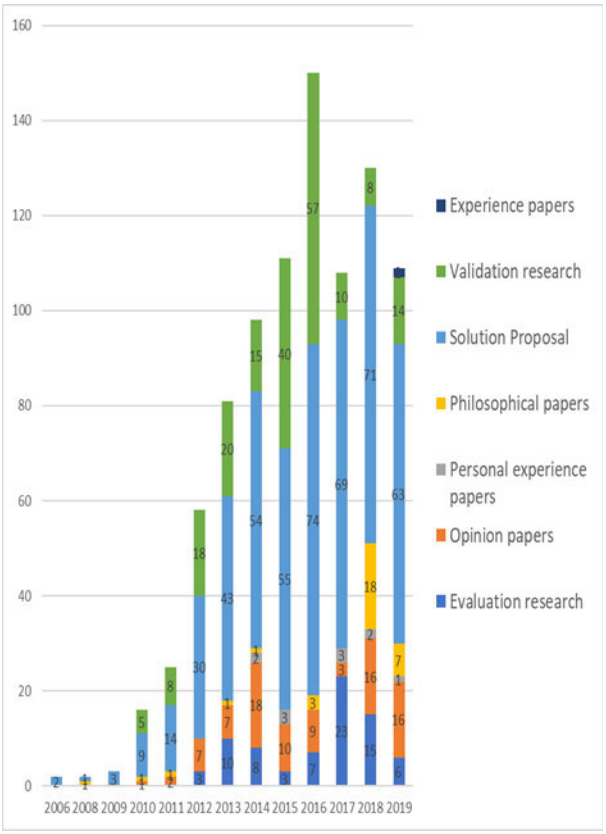

FIGURE 13. Distribution of publications according to type.

began with two scientific articles and reached a total of 62 in 2019. The ''data harmonization'', ''data selection'' and ''data curation'' phases had a progression rate of less than 3%. Data publishing is most and most relevant from 2012 and it is continuously growing until 2019.

F. FACET 6. TYPE OF RESEARCH

Figure 12 shows the distribution of publications according to research type (facet 6). We can see that the most frequent type is ''solution proposal'' with 55%, accounting for a majority of publications, followed by ''validation research'' with 22%, a much lower percentage but an important one nonetheless.

Figure 13 shows the progress of publications from 2006 to 2019 with respect to facet 6 (type). ''Evaluation'' papers tend to grow in recent years, as the field of open data has established itself. However, the number ''validation'' papers are decreasing, while papers on ''solution proposal'' remain rather stable. It is worth noting that ''philosophical'' papers are not relevant until 2018.

G. COMBINING FACETS. - THE SYSTEMATIC Map(S)

For a complete analysis of the systematic mapping results, facets were combined. Specifically, we considered four facets, namely domain, topic, life-cycle, and type of research; while the remainder facets (publication venues and impact) were not combined. Publication venues facet is not considered because of its dispersion, i.e., there are many venues with only 1 publication, so the bubble chart would not have sense, and impact facet was not considered because of the concentration, i.e., 70% of publications have an international impact and then a chart is not relevant. The results are presented in six bubble charts, each representing different possible facet combinations.

Figure 14 is the result of combining the domain facet with the data life-cycle phase facet. The majority of ''data exploitation'' publications (130) and ''data exploration'' publications (78) were related to the domain of ''infomediaries''. We also

observed that publications including a ''data exploitation'' phase were distributed over almost every domain.

Figure 15 combines the domain facet with the topic facet. The largest number of ''semantic web'' publications (87) were related to the domain of ''infomediaries''. We also found that the main domain in ''government'' publications was ''infomediaries'' (100). There is also an important number of publications on ''semantic web'' topic from the ''culture'' domain (27). ''Economy'' domain papers are mostly classified as ''government'' or ''entrepreneurship'' topics (17, and 19 papers respectively). Interestingly, a considerable number of papers (21) on ''geospatial'' domain are classified as ''software engineering'' topic. Finally, most of papers on ''transport'' domain are classified either as ''software engineering'' (11) or ''IoT'' (10).

In Figure 16, the domain facet was combined with research type. ''Solution proposals'' and ''validation research'' publications were distributed over almost all domains. The greatest amount of publications were classified as ''solution proposals'' and were related to the domain of ''infomediaries'' (212). Publications classified as ''validation research'' (74) were mainly related to the domain of ''infomediaries'', as well.

Figure 17 combines the topic and data life-cycle facets. The largest number of publications classified as ''software engineering'' (95) were related to the ''data exploitation'' phase, and most publications classified as ''semantic web'' were related either to the ''data exploitation'' (57) or ''interlinking'' (62) phase. Regarding ''government'' topic, most publications were classified as ''creation'' (42) or ''exploitation'' (46). It is surprising that there were few publications on ''harmonization'' phase classified as ''government''. Also, papers classified as ''visualization'' topic are mainly focused on ''exploitation'' and ''exploration'' phases. Most papers classified as ''information retrieval'' are related to ''creation'' phase (39). The ''semantic web'' topic accounted for a total of 193 publications; ''software engineering'' for 172, and ''government'' for 164. In the data life-cycle, the most relevant topics were: ''data exploitation'' (355 publications) and ''data exploration'' (143).

Figure 18 combines the topic facet with research type. The largest number of publications related to the ''semantic web'' topic, accounting for a total of 124, and regarding ''software engineering'' (111), all were classified as ''solution proposals''. The most frequent type of research was ''solution proposal'' with 488.

Figure 19 combines the data life-cycle facet with research type. The largest number of publications belonged to the ''exploitation'' phase (194) and they were classified as ''solution proposal''. In addition, also a number of publications in the ''data exploitation'' phase (96) were classified as ''validation research''. Most ''opinion papers'' are classified on the ''creation phase'' of the open data life-cycle (35). Evaluation papers were mainly focused on ''exploitation'' (36) and ''exploration'' (14) phases. A total of 354 publications were related to ''exploitation'', followed by 143 publications related to ''exploration''. ''Solution proposals'' (487) were predominant.

IV. DISCUSSION

In the previous section, the data analysis objectively described the number of publications per facet, based on the applied classification and taking into account evolution over time. In this section, we discuss the results in order to answer the research questions and we identify research areas, gaps, trends and open research topics.

RQ1 What are the publication venues in which open data research has been published? - The distribution of publications over the venues is shown in Figure 2 and Figure 3 for journals and conferences respectively. After analysing the results, we found that venues revealed two important communities in open data research: (i) one dedicated to Web topics with special emphasis on the Semantic Web (as well as related topics) such as the Semantic Web journal or the WWW conference, and (ii) another dedicated to e-government and its relationship with open data, i.e., the intersection between information technology and government publications such as Government Information Quarterly or conferences such as DG.O. On the other hand, there were other interesting venues with emerging communities, e.g. data engineering, information systems or artificial intelligence. These novel communities will enrich the open data arena beyond Semantic Web technologies, such as the integration of open data for Business Intelligence [22]. Therefore, the final remarkable fact is that open data research spans various research communities, which reflects the multidisciplinary nature and the variety of challenges involved in managing open data from a technological perspective. Consequently, it could be interesting to encourage multidisciplinary research, starting in the intersection of the two most important communities in open data research, as previously stated: Semantic Web and E-Government.

RQ2.-What impact have these studies had? - Figure 5 shows that 67% of the classified publications had an international impact, followed by 20% of publications with a national impact, and 7% a national impact. Therefore, although open data are closely related to public institutions that have local, regional or national regulations, most publications had an international applicability. For example, researchers worked with librarians from the ZBW-Leibniz Information Centre for Economics and at international library meetings [23], where several Linked Open Data (LOD) research topics (such as data integration and schema integration, distributed data management among others) can not only be used locally but also internationally. Other example is reported in [24], in which data was taken from two relevant services provided by a regional weather service in Spain (MeteoGalicia) –the generation of climate reports with monthly descriptions of climate behaviour and the generation of meteorological predictions– that became the core of a framework to generate linguistic descriptions of the weather. Therefore, the framework could be used by any

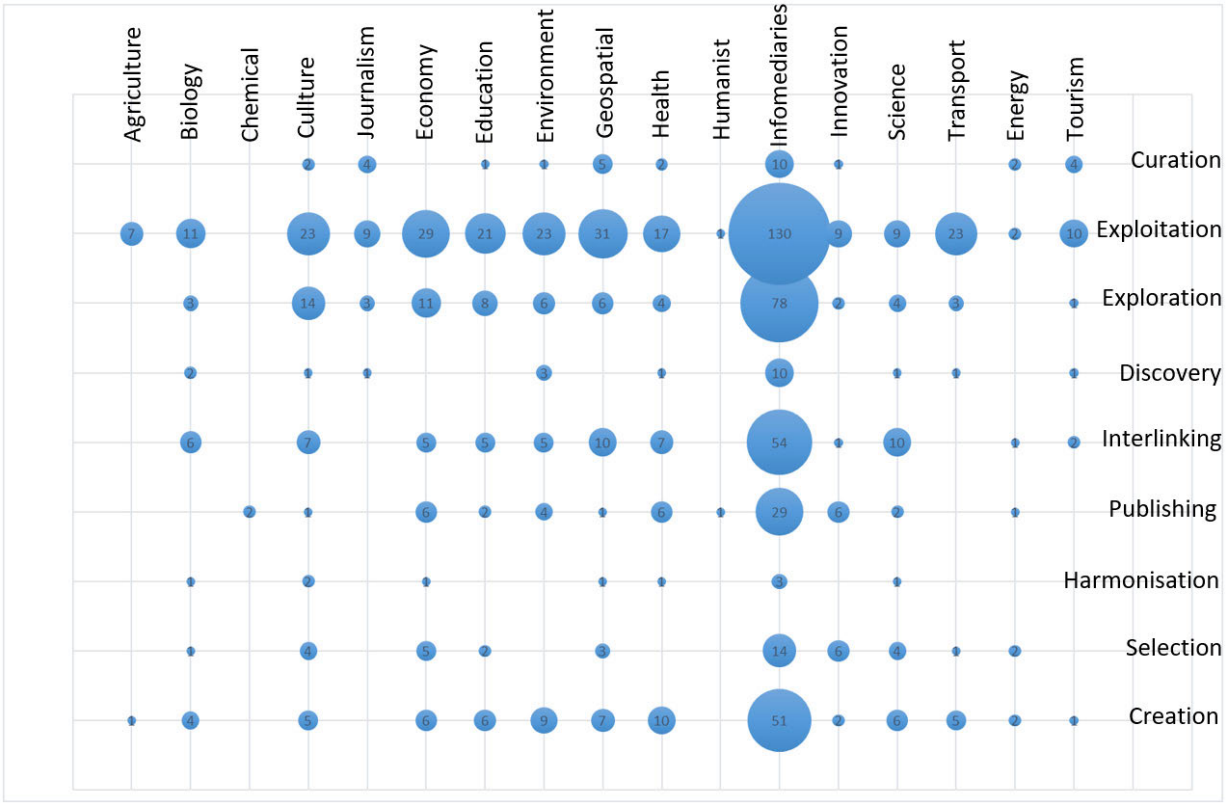

FIGURE 14. Domain & Phase of Data Cycle.

meteorological service in the world. In [25], semantic web technologies were used to understand the description of entities associated with the creation and maintenance of historical heritage in Bangladesh, but researchers can use this data in other regions around the globe.

To conclude, based on our analysis, publications on open data often refer to a specific geographical location promoted by public institutions, but research publications have a broader scope since they become a reference for other wider geographical contexts. Consequently, our results point out that research is performed locally thanks to some interesting open data local scenarios, but it can be applied worldwide.

RQ3.- What domains have been considered by the researchers? – From Figure 6, we conclude that there is an overall clear focus on the ''infomediaries'' domain (almost 43%). We put forward that most publications rely on this domain due to the importance of researching how technologies can be used for open data to benefit society (e.g., research proposals that solve problems experienced by infomediaries, facilitating the reuse of data, as well as studies on how to perform this reuse to get maximum value from data). As a matter of fact, the Open Educational Resources (OER) movement poses challenges inherent to discovering and reusing digital educational materials from highly heterogeneous and distributed digital repositories (we could highlight the work in [26], where authors presented the specifications of a data consumer oriented platform for open data, the Data-TAP, which provides an easy to use and understand interface for making educational open data friendlier to consumers, in the line proposed in [27]).

Apart from papers classified as ''infomediaries'', about 7% of papers were classified as ''culture'', ''economy'' and ''geospatial'', while 5% were classified as ''health''.

Figure 7 shows that, after increasing sharply in 2012, the number of papers related to infomediaries remained rather constant until 2017. However, number of research papers from ''infomediaries'' domain strongly decreases in 2018 (from 65 to 25 papers) and its number keep stable in 2019 (30 papers). Interestingly, other domains gain importance in those years (2018 and 2019), namely, ''journalism'' ''humanitarianism'' and ''innovation''. Those three domains deserve to be further addressed by the research community. On one hand, there are plenty of heterogeneous open data (not only structured data, but textual or multimedia data) in the humanitarian domain (such as disaster management data), which pose some relevant challenges to be solved by research community working on managing complex data. On the other hand, data journalism is related to data visualization technologies which suggests important challenges to be addressed by research community on UX (user experience), such as approach for non-expert users to analyse data or proposals for user-friendly open data management. Furthermore,

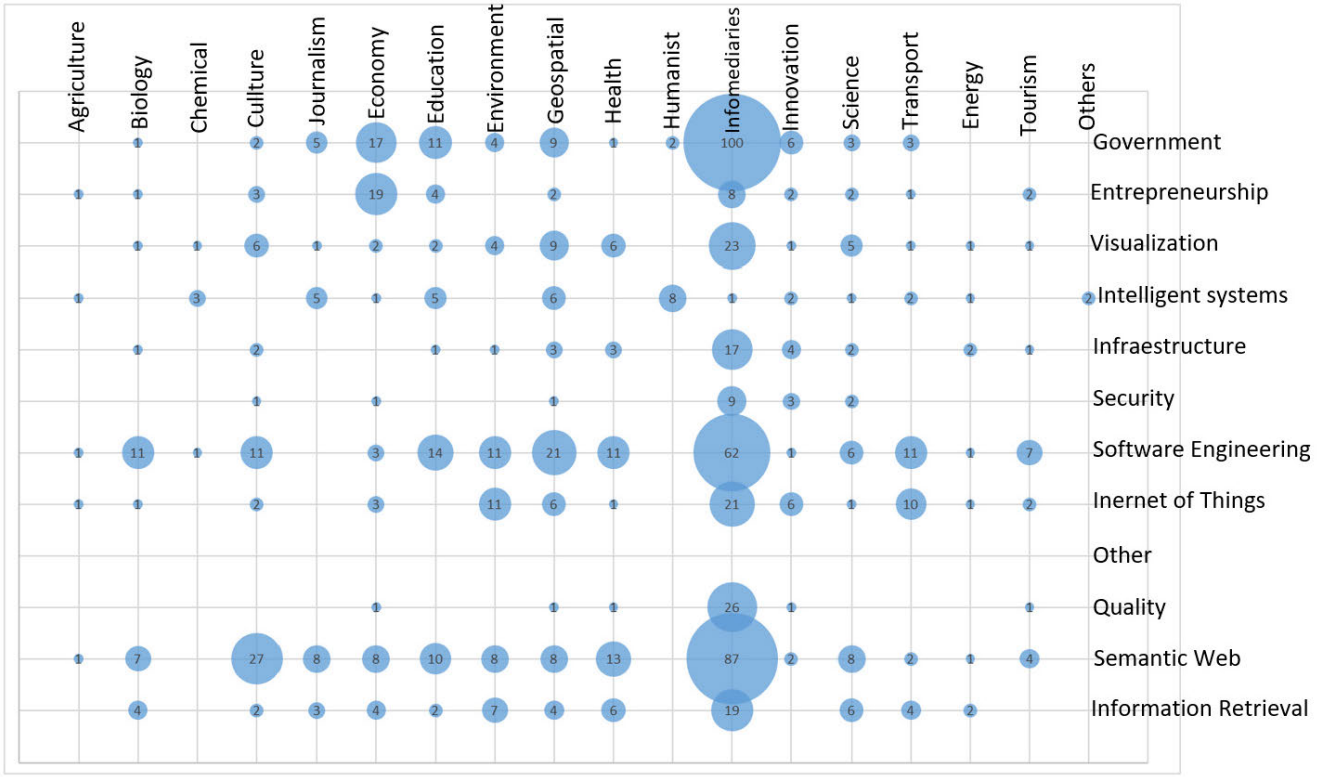

FIGURE 15. Domain and Topic.

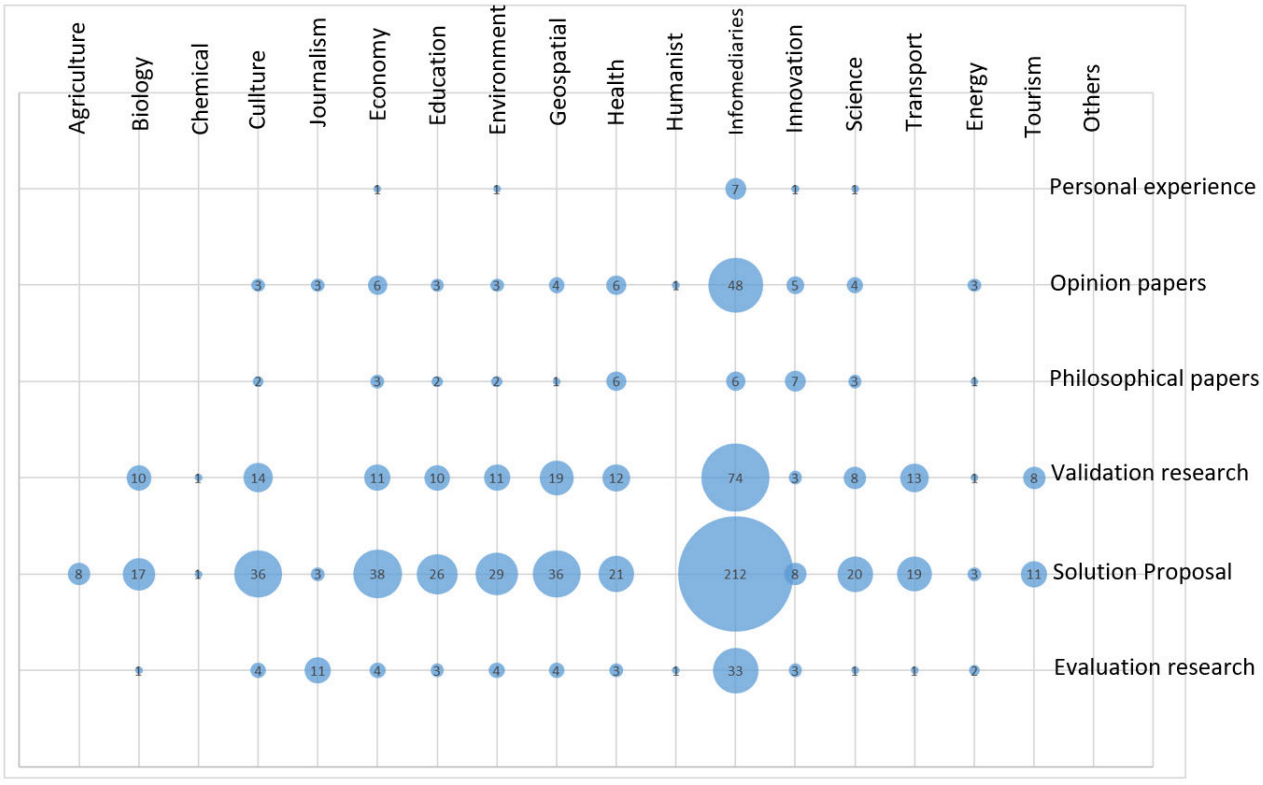

FIGURE 16. Domain and Type.

"innovation" is the second most important domain in the last two years (2018 and 2019) just after "infomediaries" domain since technological advances around open data are crucial for fostering innovation. According to Figure 14,

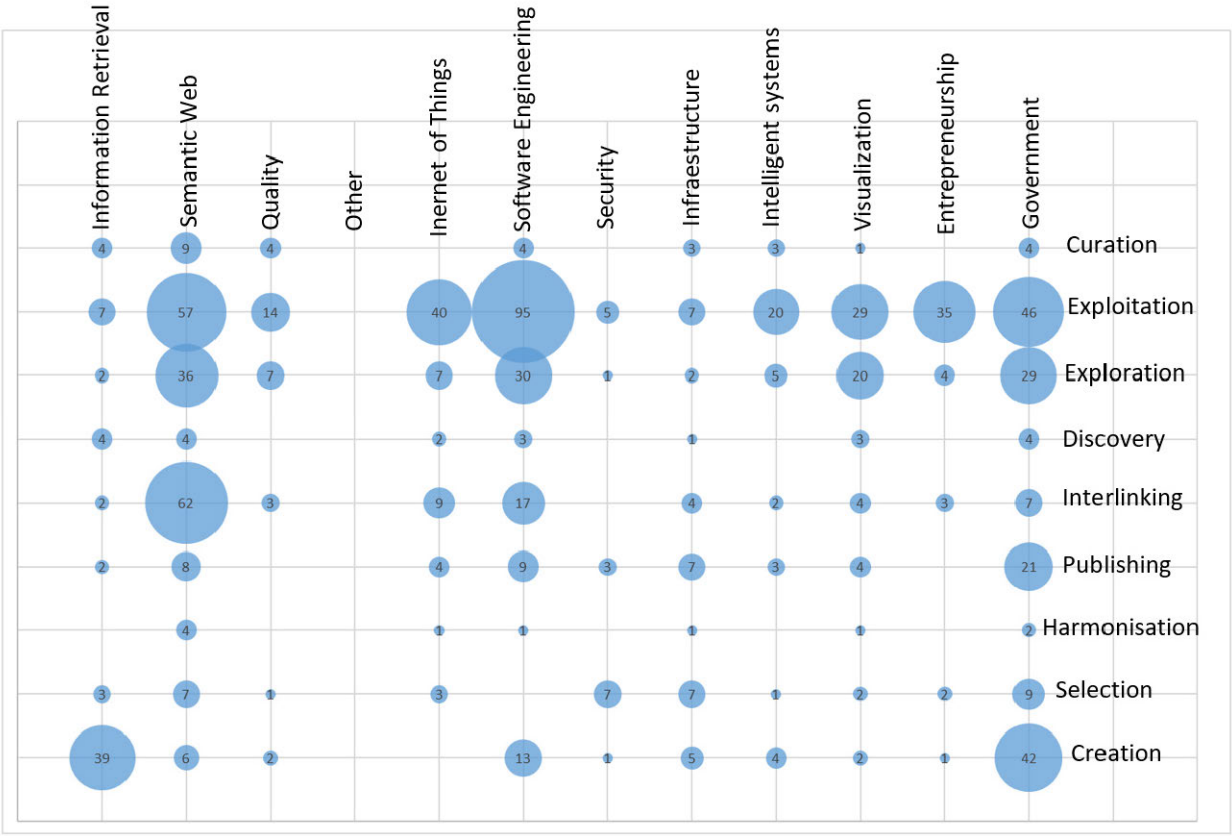

FIGURE 17. Topic and Data Life-Cycle Phase.

life-cycle phases of papers classified as ''innovation'' are mainly on ''publication'', ''selection'', or ''exploitation''; so these phases could be considered as a good starting point for further research on a more original and effective manner of using open data for a meaningful impact in society. Importantly, there appears to be room for additional research on (i) organizing open data beyond open data portals in order to facilitate searching and integration; (ii) improving how open data is using in the artificial intelligence arena; or (iii) envision how open data innovation may become into a real business enabler. Due to the relevance of the ''innovation'' domain in recent years, we have included a specific section (Section V) for a proper discussion on reviewing open data innovation projects by considering both open data publication (e.g., open data portals) and open data consumption (e.g., open data as a business enabler).

Other domains (such as transport, economics, education, environment, culture, health and geospatial) have increased significantly in the same proportion over the past seven years (from 2006 to 2009 there was no research in these areas). This growth could be due to the fact that these domains use open data to provide solutions and benefits to citizens, i.e. social challenges that require research on novel technologies to solve them.

Surprisingly, managing open data for tourism received little attention from the research community, since very few papers were classified in the ''tourism'' domain each year. We can conclude that considering technologies for better managing open data in the tourism domain is an open research issue that need to be furthered considered (e.g. related to the smart tourism destination concept [21], which considers the smart city scenario from a touristic point of view).

Figure 14 shows the relationship between the domain and the life-cycle data phase, with 130 publications that relate the exploitation of open data to infomediaries, thus revealing the need to use technologies for supporting open data to solve real problems in society. Actually, ''exploitation'' phase is the most important phase for almost each domain: (i) in the domain of ''culture'', developments stem from the policy of opening data from bibliographic sources that are reused in museum applications; (ii) the ''economy'' domain is growing, and this rise is attributed to entrepreneurship use of data; (iii) the ''geospatial'' domain has developed because of growing use in maritime, safety and health applications. However, there are some exceptions, such as papers classified as ''science'' domain that are mainly classified as being in the ''interlinking'' phase, due to the great importance of open science as manner of provide an accessible knowledge that

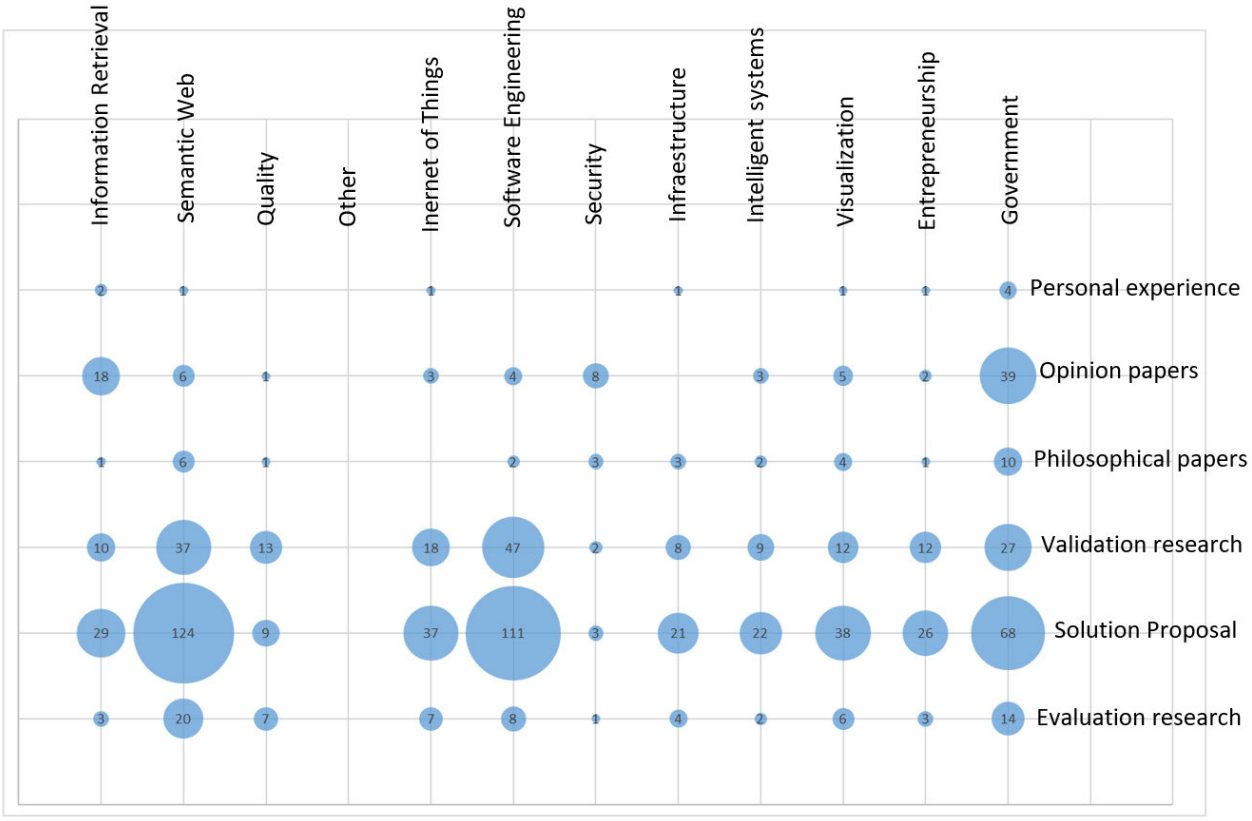

FIGURE 18. Topic and Type.

can be shared and developed through collaborative networks of researchers.

In Figure 16 (bubble chart where domain facet was combined with research type facet), papers on ''solution proposals'' (488 out of 893) and ''validation'' (195 out of 893) are distributed over all domains, but clearly mostly applies to ''infomediaries''.

RQ4.- Which phases in the data publication life-cycle have been considered in the studies? - Figure 10 shows the contributions of papers classified by data life-cycle: 40% corresponded to ''data exploitation'', 16% to ''data exploration'', and together, these two phases of the life-cycle represented 56% of the published papers. Figure 11 shows that since 2010, both phases grew, responding to the need to gain insights from data in real-life applications. This process has been led by infomediary community requirements regarding data reuse (which agrees with our previous discussion about the most important research domain on open data, namely ''infomediaries''). The first phases of the data life-cycle are ''publication'', ''creation'', ''interlinking'' and ''harmonization''. Specifically, figure 10 shows that 13% of papers are dedicated to ''data creation'' and 7% to ''data publication''. This contrasts with 40% of papers related to ''data exploitation'', 16% to ''data exploration'' and 13% to ''data interlinking''. This leads us to conclude that more effort is required in the research community on data life-cycle phases relating to how data is published. To date, published data are not yet of sufficient quality as they are being published and exploited without going through life-cycles that can add quality. Further research on ''data curation'' for improving open data quality is thus required. Therefore, the ''data publication'' phase of the life-cycle needs to be emphasised in open data research. Research covering all the data life-cycle phases proposed by [1] is highly relevant, because a preparation process prior to publication guarantees better reuse.

Phases related to data publication are crucial due to the inherent technical complex nature of data publication together with the rapid process of data digitization, which lead to problems of format standardization and connections to heterogeneous data sources [29]–[32]. Those contributions are important for understanding that it is necessary to work all stages proposed in the data life-cycle with special emphasis on the phases that prepare the data for publication. In this sense, proposals exist to improve publication time using methodologies based on automated procedures to prepare the data for publication [28]. However, also worthy of note, research on open data focused mainly on ''exploitation'' and ''exploration'' (i.e., on the life-cycle phases directly related to data consumption).

Importantly, our results show that ''data interlinking'' represented 13% of all reviewed documents, with special emphasis on papers classified as ''science'' (as we explained above

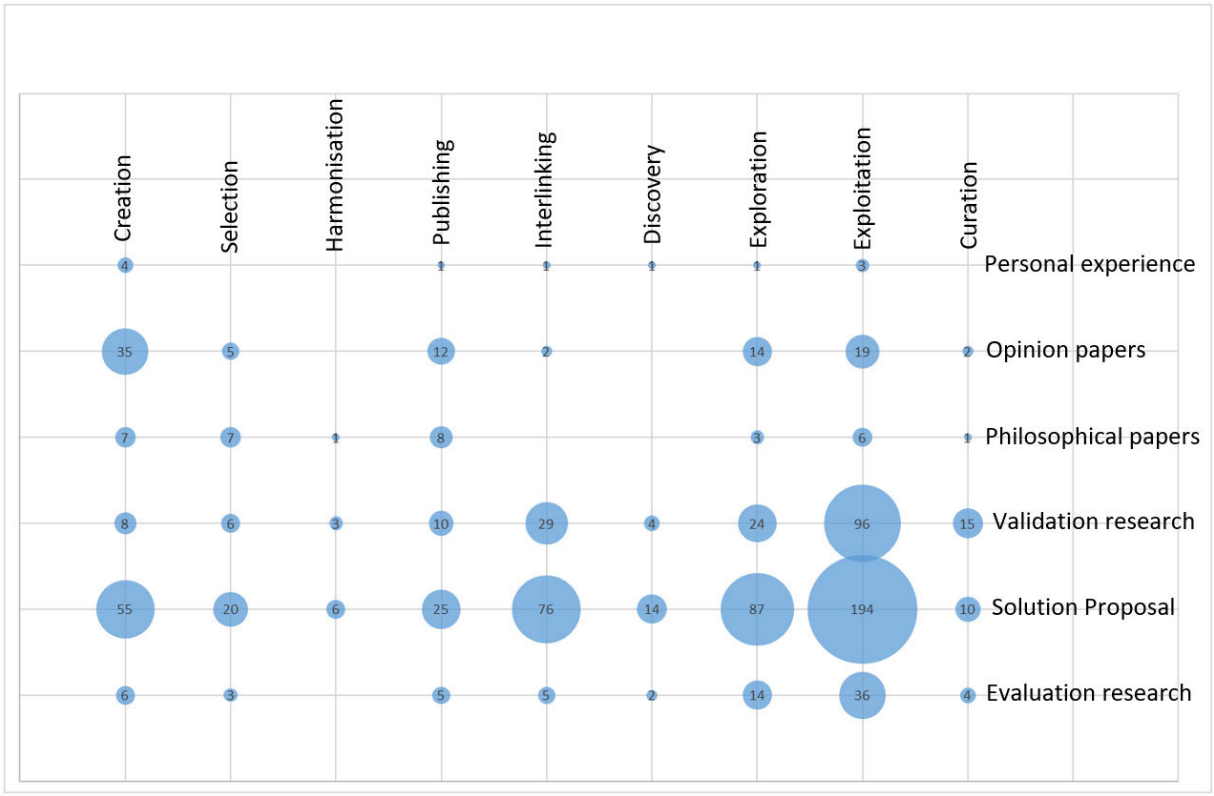

FIGURE 19. Data Life-Cycle and Research Type.

in RQ3). This data phase has grown steadily over time due to the rise of the semantic web as a means of producing those processes required for a powerful publication of open data. Finally, ''data harmonization'', ''data selection'' and ''data curation'' phases had a development rate of less than 3%, while ''data discovery'' and ''data publishing'' grew at a slower rate.

Figure 19 shows the data lifecycle phases according to type of publication. The majority of papers (194) refers to ''solution proposals'' within the ''data exploitation'' phase, confirming that the research community is focusing on presenting solutions for reusing data.

RQ5.- What types of research are performed? - Figure 12 shows that 55% of the research papers consisted in ''solution proposals'', followed by ''validation research'' with 22%, ''opinion papers'' with 10%, ''evaluation research'' with 8%. Less frequent types of research were ''experience paper'' with 1%. Regarding, the progress of publications from 2006 to 2019, number of ''philosophical'' papers burst in 2018.

Figure 18 relates the research topic to the research type. Once the data was analysed, a large number of publications was classified as ''solution proposal'' and ''validation research''. Many solution proposals were oriented towards the ''semantic web'' topic (124), responding to the need to relate open data to create more efficient applications that integrate data from a disparate of interlinked sources. The proposed solutions originated in response of infomediaries requirements, i.e. ''infomediaries'' domain (212) as shown in Figure 16. The infomediaries provided solutions based on open data but according to our mapping, attention is increasingly paying on data quality. According to [28], open data quality is one of the main threats to reach the objectives of the open data movement.

Very few papers were registered as opinion papers or philosophical papers as efforts are being directed towards practical solutions aimed at being validated and evaluated. An example of open data evaluation is the proposal of analysis of clinical trial reports, where research evaluation plays an important role before the solution is widely adopted [33]. Another example of a semantic web solution proposal is that of a unified framework for converting open legacy data into graphical images and into a machine-readable and reusable format using crowdsourcing [34] to allow elder people to access open data. Meanwhile, [35] has addressed the information needs of people with mobility impairments relating to walking distances and the existence of ramps, steps, seats, etc.

These examples allow us to understand that solution proposals are important and closely related to the development of the open data concept. Furthermore, the amount of evaluation and validation papers increased as from 2010, which shows the open data research community has reached a certain

stage of maturity. Nevertheless, it would be desirable that researchers put the same effort into conducting evaluations and validations to support these solution proposals. Therefore, research publications that not only describe solution proposals related to open data, but conduct evaluation and validation experiments, are still highly encouraged. A larger number of publications on evaluation and validation would entail that the research artifacts produced by open data researchers become mature enough faster for potential transfer to the industrial field.

RQ6.- What topics were addressed by research? - Figure 8 shows that papers about ''semantic web'' topic accounts for 24% of the research. The second most popular topic is ''software engineering'' with 22% followed by ''government'' with 16%. Significantly, the topic of ''information retrieval'' accounted for 8%. while ''Internet of Things'' represents 8% of the publications.

Figure 15 relates domain and topic. Topics on ''semantic web'' (87), ''government'' (100) and ''software engineering'' (62) accounted together for 28% of total publications. The remarkable development of ''semantic web'' and ''software engineering'' publications is based on the tendency to exploit and explore data, in different domains discussed in the research question 3 (RQ3) explained above. The large number of publications relating to ''government'' reveals the need for a legal framework with policies and regulations to manage open data but from a technological perspective. Therefore, multidisciplinary research on legal and technological open data issues is much needed to establish policies that facilitate the development of open data and generate secure and reliable technological services using open data.

''Semantic web'' is a crucial research topic for open data due to the need for data schemas to exchange data. This process enables to automatically generate information to develop applications and facilitate platforms for example in the areas of health [37], or the visualization of environmental issues, mapping of utility management, evaluation of political lobbying, social benefits, closing the digital divide, biology and others [38].

The development of ''software engineering'' research applied to various areas is notable in applications for using open data in teaching or development of APIs for tourism applications [39], analysis of data generated by smartphones, open data collection to present useful information to citizens for public events [40], [41], [21], [42]–[44].

''Software engineering'' accounted for 172 articles and it was the second most frequent topic. This can be explained by users' need to reuse data. The biggest number of contributions was in 2015, with 18% of the total. However, this result contrasts with the development of entrepreneurial projects, since software applications were aimed to reuse services, but not entrepreneurship. We must remember that entrepreneurship does not necessarily rely on research, which also may explain why there were few scientific studies on entrepreneurship and open data. However, we deeply believe that it is a research gap to study how technologies can be used to encourage open data for ''entrepreneurship'' topic.

Finally, research and development on the topic of ''security'' has been limited despite the fact that the topic is relevant from the open data viewpoint. The ''IoT'' topic grew steadily until 2019, responding to the development of smart cities and sensors.

V. BEYOND SYSTEMATIC MAPPING: OPEN DATA INNOVATION PROJECTS

In addition to a quantitative analysis of the research regarding open data over the last years, presented in the previous sections, we now go beyond the systematic mapping study and present a review of open data innovation projects. We focus on two phases of the data life-cycle: publication (e.g., open data portals) and consumption (e.g., open data as business enabler) of open data.

A. OPEN DATA PORTALS AND DATASET SEARCH

The first critical step of the data life-cycle when considering external agents that may consume the data is its publishing. Typically, this is done by setting up an open data portal (ODP), where the particular administrative division, or public organization that produces or collects the data makes it accessible to the public [45], proposes a classification of ODPs according to the number of functionalities they provide, ordered by the amount of effort to set them up that is required.

- A ''dataset registry'', a simple list of links towards datasets, not necessarily hosted in the portal. A registry answers the question ''Who has which (open) dataset and where can I find it?''
- A ''metadata provider'', a dataset registry that also holds metadata about datasets, e.g., licensing, spatio-temporal context, update frequency.
- A ''Co-creation platform'', is a metadata provider that includes tools of participation for citizens and data consumers to actuate on datasets: generate ideas, raising issues, contribute with re-use examples, and/or participate in discussions.
- A ''Data publishing platform'', a Co-creation platform that enables multiple data providers to publish their own datasets. It also supports the interlinking phase among the hosted datasets
- A ''Common data hub'', a data publishing platform that also supports the implementation of further phases of the data life-cycle, enabling data publishers to implement their own data cycles.

Open source and commercial software systems like CKAN, Socrata, and OpenDataSoft enable open data providers to easily set up a portal between levels 3 and 4. To provide a unified view that facilitates search at national and supranational levels from regional and local portals, meta-portals crawl and index datasets into a central location where further interlinking can also be performed. An

example formalization of such a model is the MODA (Middleware for Open Data) framework, described in [46]. Many national level portals proceed this way, aggregating data from regional portals and from government agencies. The Europeana project [47], provides a single access point to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe. The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries. Information regarding the provision of data and the benefits of re-using data is also included.

Once datasets are published, the next step facing the user is to provide an appropriate search functionality. Dataset search is a relatively new field of research that lies in the intersection of information retrieval, databases, semantic web, and enterprise data management. The first effort was the International Open Government Dataset Search [48], a faceted browsing interface for searching over more than one million open government datasets from around the world. Challenges for Dataset Search were first outlined in [49] in the context of scientific data, proposing a Crawl-Read-Extract process similar to document search engines. More recently, there have been efforts in understanding the subtleties of dataset search from a user perspective, either by interviewing practitioners [50] or by analyzing query logs and data requests [51]. Current research is focusing on using these insights to develop machine learning models for ranking datasets in a portal according to a query [52], [53].

Dataset search [54], including but not limited to open datasets, has been recently identified as a research field on its own on its own right, that broadly encompasses frameworks, methods and tools that help match a user data need against a collection of datasets, recognizing datasets as interesting entities to themselves with some properties shared with documents, tuples and webpages, and some unique to them. The development of benchmarks for evaluating dataset search algorithms has been identified as the most immediate open research problem.

B. OPEN DATA INNOVATION AS BUSINESS ENABLER

Open Data has been proven as a great tool for increasing transparency and empower citizens, moreover, several works have identified the potential of open data as a catalyzer for innovation, therefore, enabling the creation of value and services that ultimately benefit citizens. In the context of Open Government Data (OGD), the work of [55] identifies 4 mechanisms for mixed social-economic value generation: Transparency (that improves visibility), participation (engagement with all stakeholders), efficiency (cost and time reduction) and innovation (generation of new ideas), highlighting data openness as an enabler of both the generation and appropriation of value. [53] surveyed 138 Swedish IT-entrepreneurs, finding that access to public open data is considered very important for many of them; 43% find open data essential for the realization of their business plan and 82% claim that access would support and strengthen the business plan. Entrepreneurs also showed interest in, and willingness to pay for, public sector information data to support or test other business models. Reference [56] analyzed data from 500 US-based firms that use open government data in their business model and proposed a taxonomy of business model archetypes: Enablers of collection, management, and disclosure of public data; Facilitators that support or accelerate the access and exchange of data between the supply side and the demand-side; and Integrators, that make use of open government data by combining it with internal data or other types of proprietary data in order to augment its business capabilities. On the other hand, [57] identifies as a perceived blocker the lack of competitive advantage, as open data is accessible to everyone (including competitors), their study suggests that the generation of competitive advantage with open data requires a company to have in-house capabilities and resources for open data use.

Concerning innovation activities, two recent EU-funded innovation actions looked at how to unleash the potential of Open Data as a business enabler, by providing SMEs with the capabilities needed to process open data, plus general business support, inline with the recommendations in [58]. We briefly describe below how they worked and their results.

The Future INternet and Open Data EXpansion (FINODEX) ran from mid 2014 to 2016. It was aimed at the promotion and support of innovative ICT services re-using open data, using as technical anchor the FI-WARE platform. Two open calls were organized for SMEs to submit their product and service ideas. Selected SMEs were provided with funding, support, tailored training, networking opportunities, and connection with investors. The time and funding of the SMEs inside the incubator was dictated using a ''funnel'' approach. Initially, all SMEs received a certain amount of funding to deliver an initial milestone, that is evaluated to decide which SMEs proceed to the next phase, where they received a further amount of funding against a new milestone, and so on up to 4 milestones have been achieved. The highest amounts of direct funding: €170.000, €135.000 and €115.000, was assigned to the first, second and third project on both rounds of acceleration.

The Open Data Incubator for Europe (ODINE) ran from mid 2015 to mid 2017, aimed at incubating business ideas centered around open data. Compared to FINODEX, it did not use the funnel approach to incubation, but a flat 6 months incubation period for all selected companies, that where selected in a rolling open call of 8 iterations and did not force SMEs to use the FIWARE stack, and with a maximum funding of €100.000 per SME. ODINE incubated 57 companies that created 278 new jobs created and €23.7M on sum of sales and investments. ODINE's impact on the growth perspectives of the funded companies was relevant, resulting in an estimated €110 M of cumulative revenues in the period 2016-2020, plus 784 jobs created. An independent impact assessment study on ODINE prepared by IDC [59] found that most funded companies where young, and played the role of ''experimenters'', that is, combining several

open data sources to improve their product and services. The assessment also observed a positive correlation between the level or maturity at country level of the Open Data market and the number of ODINE successful applicants by country, suggesting that a rich open data environment provides favorable conditions for innovators in this field.

VI. CONCLUSION

A systematic mapping study was carried out in this paper in order to assemble, classify, and analyse all research on open data from a technological perspective performed between 2006 and 2019 (both inclusive) by the scientific community with the aim of: (i) providing a consolidated overview of the research field, and (ii) identifying, among others, well-established topics, trends, and open research issues. Our study revealed several interesting facts:

- Most research on open data from a technical perspective came from the IEEE and ACM scientific repositories.
- The number of papers on open data published before 2009 was not significant. Incidentally, it was in 2009 that the United States established an Open Government directive under the Obama administration.
- Publications from 2006 to 2009 were incipient (in fact there were no publications on open data in 2007).
- The ''sematic web'', ''software engineering'' and ''government'' were some of the most important topics addressed in the research. We believe this is to (i) Semantic Web technologies are intrinsically related to open data reusability, (ii) software development is required to solve technical problems relating to data openness, and (iii) legislation and standardization is needed for society to more profusely use open data.
- The topics of ''Internet of Things'' and ''quality'' are not yet strongly developed, but given current technological evolutions, these topics may become highly significant in the near future with the implementation of smart cities and the importance of sufficient data quality to support informed decision-making.
- ''Infomediaries'' was the most developed domain in the publications. The other domains represented under 8% each; this suggests that publications were directed especially to information channels. Nevertheless the ''geospatial'', ''health'', and ''culture'' domains proved to be an ongoing object of research and these studies continue to grow.
- When analysing the phases, ''exploitation'' and ''exploration'' were the most frequent, showing the community's need for practical applications of open data. Surprisingly, phases relating to data publication were underrepresented. These phases have received insufficient attention and should be researched further in the future.
- In the same way, the ''solution proposal'' and ''validation research'' types were the most frequent, showing that the field is reaching maturity.
- Concerning impact, international publications were the most frequent because software developments and semantic web publications are internationally applicable and accepted.
- In addition to the analysis provided by our systematic mapping study, and due to the fact that ''innovation'' is one of the most relevant domains in the last two years (2018 and 2019), a survey of research insights regarding open data innovation projects was performed, with special emphasis on publication (e.g., open data portals) and consumption (e.g., open data as business enabler) of open data.
- In 2019, research interest in open data from a technological perspective overall decreased. This fact may indicate that research is stabilizing, since the open data research hype is somewhat over. This could be related to the Gartner Hype Curve, thus indicating that open data research will reach maturity, e.g. the amount and quality of research performed can be considered to be consolidating as the number of solutions proposed (research contribution type) decreased in 2019, while at the same time, more validation and evaluation studies are being conducted.

Finally, we hereby encourage the open data research community to address the under-researched topics and fill the gaps in open data research from a technological perspective emphasising the need to create multidisciplinary research teams.

REFERENCES

[1] J. Attard, F. Orlandi, S. Scerri, and S. Auer, ''A systematic review of open government data initiatives,'' *Government Inf. Quart.*, vol. 32, no. 4, pp. 399–418, Oct. 2015.
[2] T. W. House, ''The Obama administration's commitment to open government: A status report,'' White House, Washington, DC, USA, Tech. Rep., 2011.
[3] F. Gonzalez-Zapata and R. Heeks, ''The multiple meanings of open government data: Understanding different stakeholders and their perspectives,'' *Government Inf. Quart.*, vol. 32, no. 4, pp. 441–452, Oct. 2015.
[4] Open Data Charter, ''International open data charter,'' Open Government Partnership, Washington, DC, USA, Tech. Rep., Sep. 2015, p. 8.
[5] M. Janssen and J. van den Hoven, ''Big and open linked data (BOLD) in government: A challenge to transparency and privacy?'' *Government Inf. Quart.*, vol. 32, no. 4, pp. 363–368, Oct. 2015.
[6] K. O'Hara, ''Transparency, open data and trust in government,'' in *Proc. 3rd Annu. ACM Web Sci. Conf. (WebSci)*, vol. 2, 2012, pp. 223–232.
[7] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, ''Systematic mapping studies in software engineering,'' in *Proc. 12th Int. Conf. Eval. Assess. Softw. Eng.*, vol. 17, 2008, p. 10.
[8] W. Afzal, R. Torkar, and R. Feldt, ''A systematic review of search-based testing for non-functional system properties,'' *Inf. Softw. Technol.*, vol. 51, no. 6, pp. 957–976, Jun. 2009.
[9] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman, ''Systematic literature reviews in software engineering—A tertiary study,'' *Inf. Softw. Technol.*, vol. 52, no. 8, pp. 792–805, 2010.
[10] A. Sutton, M. Clowes, L. Preston, and A. Booth, ''Meeting the review family: Exploring review types and associated information retrieval requirements,'' *Health Inf. Libraries J.*, vol. 36, no. 3, pp. 202–222, Sep. 2019.
[11] S. Casteleyn, I. Garrig'os, and J.-N. Maz'on, ''Ten years of rich Internet applications: A systematic mapping study, and beyond,'' *ACM Trans. Web*, vol. 8, no. 3, pp. 1–46, Jun. 2014.
[12] (2013). *G8 Open Data Charter, G8 Lough Erne*. Accessed: Jun. 2013. [Online]. Available: https://opendatacharter.net

[13] J. Hagel, III, and J. F. Rayport "The coming battle for customer information," *Harvard Bus. Rev.*, vol. 3, p. 64, Jul. 1997.

[14] R. Wieringa, N. Maiden, N. Mead, and C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion," *Requirements Eng.*, vol. 11, no. 1, pp. 102–107, Mar. 2006.

[15] J. Rathbone, T. Hoffmann, and P. Glasziou, "Faster title and abstract screening? Evaluating abstrackr, a semi-automated online screening program for systematic reviewers," *Systematic Rev.*, vol. 4, no. 1, pp. 1–7, Dec. 2015.

[16] B. E. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney, M. Macleod, R. R. Shah, and K. Thayer, "SWIFT-review: A text-mining workbench for systematic review," *Systematic Rev.*, vol. 5, no. 1, p. 87, Dec. 2016.

[17] Z. Cai and W. Zhu, "Feature selection for multi-label classification using neighborhood preservation," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 1, pp. 320–330, Jan. 2018, doi: 10.1109/JAS.2017.7510781.

[18] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 703–715, May 2019, doi: 10.1109/JAS.2019.1911447.

[19] S. Barney, K. Petersen, M. Svahnberg, A. Aurum, and H. Barney, "Software quality trade-offs: A systematic map," *Inf. Softw. Technol.*, vol. 54, no. 7, pp. 651–662, Jul. 2012.

[20] E. Engström, P. Runeson, and M. Skoglund, "A systematic review on regression test selection techniques," in *Proc. ACM Int. Conf.*, 2010, vol. 53, no. 1, pp. 14–40.

[21] J. A. Ivars-Baidal, M. A. Celdrán-Bernabeu, J.-N. Mazón, and Á. F. Perles-Ivars, "Smart destinations and the evolution of ICTs: A new scenario for destination management?" *Current Issues Tourism*, vol. 22, no. 13, pp. 1581–1600, Aug. 2019.

[22] M. Love, C. Boisvert, E. Uruchurtu, and I. Ibbotson, "Nifty with data: Can a business intelligence analysis sourced from open data form a nifty assignment?" in *Proc. ACM Conf. Innov. Technol. Comput. Sci. Edu.*, Jul. 2016, pp. 344–349.

[23] A. Latif, A. Scherp, and K. Tochtermann, "LOD for library science: Benefits of applying linked open data in the digital library setting," *KI - Künstliche Intelligenz*, vol. 30, no. 2, pp. 149–157, Jun. 2016.

[24] A. Ramos-Soto, A. Bugarín, S. Barro, and F. Díaz-Hermida, "Automatic linguistic descriptions of meteorological data a soft computing approach for converting open data to open information," in *Proc. Iberian Conf. Inf. Syst. Technol.*, 2013, pp. 1–6.

[25] S. Chakraborty, M. H. Hafizur Rahman, and M. H. Seddiqui, "Linked open data representation of historical heritage of bangladesh," in *Proc. 16th Int. Conf. Comput. Inf. Technol.*, Mar. 2014, pp. 242–248.

[26] E. Piedra, N. Chicaiza, J. Lopez, and J. Tovar Caro, "Towards a learning analytics approach for supporting discovery and reuse of OER an approach based on social networks analysis and linked open data," in *Proc. IEEE Global Eng. Educ. Conf.*, Mar. 2015, pp. 978–988.

[27] C. Millette and P. Hosein, "A consumer focused open data platform," in *Proc. 3rd MEC Int. Conf. Big Data Smart City (ICBDSC)*, Mar. 2016, pp. 101–106.

[28] J. N. Rouder, "The what, why, and how of born-open data," *Behav. Res. Methods*, vol. 48, no. 3, pp. 1062–1069, Sep. 2016.

[29] A. O. Erkimbaev, V. Y. Zitserman, G. A. Kobzev, V. A. Serebrjakov, and K. B. Teymurazov, "Publishing scientific data as linked open data," *Sci. Tech. Inf. Process.*, vol. 40, no. 4, pp. 253–263, Oct. 2013.

[30] A. Callahan, J. Cruz-Toledo, and M. Dumontier, "Ontology-based querying with Bio2RDF's linked open data," *J. Biomed. Semantics*, vol. 4, no. 1, p. S1, 2013.

[31] S. O'Riain, E. Curry, and A. Harth, "XBRL and open data for global financial ecosystems: A linked data approach," *Int. J. Accounting Inf. Syst.*, vol. 13, no. 2, pp. 141–162, Jun. 2012.

[32] D. S. Sayogo and T. A. Pardo, "Exploring the motive for data publication in open data initiative: Linking intention to action," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2011, pp. 2623–2632.

[33] P. Ciancarini, F. Poggi, and D. Russo, "Big data quality: A roadmap for open data," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2016, pp. 210–215.

[34] P. Doshi and T. Jefferson, "Open data 5 years on: A case series of 12 freedom of information requests for regulatory data to the European medicines agency," *Trials*, vol. 17, no. 1, p. 78, Dec. 2016.

[35] S. Oyama, Y. Baba, I. Ohmukai, H. Dokoshi, and H. Kashima, "Crowdsourcing chart digitizer: Task design and quality control for making legacy open data machine-readable," *Int. J. Data Sci. Analytics*, vol. 2, nos. 1–2, pp. 45–60, Dec. 2016.

[36] N. B. Hounsell, B. P. Shrestha, M. McDonald, and A. Wong, "Open data and the needs of older people for public transport information," *Transp. Res. Procedia*, vol. 14, pp. 4334–4343, Jan. 2016.

[37] H. Demski, S. Garde, and C. Hildebrand, "Open data models for smart health interconnected applications: The example of openEHR," *BMC Med. Informat. Decis. Making*, vol. 16, no. 1, p. 137, Dec. 2016.

[38] M. Kassen, "A promising phenomenon of open data: A case study of the Chicago open data project," *Government Inf. Quart.*, vol. 30, no. 4, pp. 508–513, Oct. 2013.

[39] R. L. Pereira, P. C. Sousa, R. Barata, A. Oliveira, and G. Monsieur, "CitySDK tourism API–building value around open data," *J. Internet Services Appl.*, vol. 6, no. 1, p. 24, Aug. 2015.

[40] W. Brunette, R. Sodt, R. Chaudhri, M. Goel, M. Falcone, J. Van Orden, and G. Borriello, "Open data kit sensors: A sensor integration framework for Android at the application-level," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, vol. 12, 2012, p. 351.

[41] T. Silva, V. Wuwongse, and H. N. Sharma, "Disaster mitigation and preparedness using linked open data," *J. Ambient Intell. Humanized Comput.*, vol. 4, no. 5, pp. 591–602, Oct. 2013.

[42] F. G. De Andrade and R. José, "Semantic annotation of geodata based on linked-open data," in *Proc. 7th Int. Conf. Manage. Comput. Collective Intell. Digit. EcoSyst.*, vol. 2, 2015, pp. 9–16.

[43] Y.-A. Lai, Y.-Z. Ou, J. Su, S.-H. Tsai, C.-W. Yu, and D. Cheng, "Virtual disaster management information repository and applications based on linked open data," in *Proc. 5th IEEE Int. Conf. Service-Oriented Comput. Appl. (SOCA)*, Dec. 2012, pp. 1–5.

[44] N. Kobayashi and T. Toyoda, "BioSPARQL: Ontology-based smart building of SPARQL queries for biological Linked Open Data," in *Proc. ACM Int. Conf. Proc. Ser.*, no. 1, 2012, pp. 47–49.

[45] P. Colpaert, J. Sarah, P. Mechant, E. Mannens, and R. Van de Walle, "The 5 stars of open data portals," in *Proc. 7th Int. Conf. Methodol. Technol. Tools Enabling E-Government*, 2013, pp. 61–67.

[46] X. Masip-Bruin, G.-J. Ren, R. Serral-Gracia, and M. Yannuzzi, "Unlocking the value of open data with a process-based information platform," in *Proc. IEEE 15th Conf. Bus. Informat.*, Jul. 2013, pp. 331–337.

[47] A. Isaac and B. Haslhofer, "Europeana linked open data–data.Europeana.eu," *Semantic Web*, vol. 4, no. 3, pp. 291–297, 2013.

[48] E. Rozell, J. Erickson, and J. Hendler, "From international open government dataset search to discovery: A semantic Web service approach," in *Proc. 6th Int. Conf. Theory Pract. Electron. Governance (ICEGOV)*, 2012, pp. 480–481.

[49] D. Maier, V. M. Megler, and K. Tufte, "Challenges for dataset search," in *Database Systems for Advanced Applications*. Cham, Switzerland: Springer, 2014, pp. 1–15.

[50] L. M. Koesten, E. Kacprzak, J. F. A. Tennison, and E. Simperl, "The trials and tribulations of working with structured data: -a study on information seeking behaviour," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 1277–1289.

[51] E. Kacprzak, L. Koesten, L. D. Ibáñez, T. Blount, J. Tennison, and E. Simperl, "Characterising dataset search—An analysis of search logs and data requests," *J. Web Semant.*, vol. 55, pp. 37–55, Mar. 2018.

[52] M. de Rijke, "Learning to search for datasets," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, p. 1483.

[53] L. Mlynárová, J.-P. Nap, and T. Bisseling, "The SWI/SNF chromatin-remodeling gene AtCHR12 mediates temporary growth arrest in arabidopsis thaliana upon perceiving environmental stress," *Plant J.*, vol. 51, no. 5, pp. 874–885, Jul. 2007.

[54] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. D. Ibáñez, E. Kacprzak, and P. Groth, "Dataset search: A survey," *VLDB J.*, vol. 29, no. 1, pp. 251–272, 2020.

[55] T. Jetzek, M. Avital, and N. Bjorn-Andersen, "Data-driven innovation through open government data," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. 100–120, 2014.

[56] E. Lakomaa and J. Kallberg, "Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs," *IEEE Access*, vol. 1, pp. 558–563, 2013.

[57] G. Magalhaes, C. Roseira, and L. Manley, "Business models for open government data," in *Proc. 8th Int. Conf. Theory Pract. Electron. Governance*, Oct. 2014, pp. 365–370.

[58] A. Zuiderwijk, M. Janssen, K. Poulis, and G. van de Kaa, "Open data for competitive advantage: Insights from open data use by companies," in *Proc. 16th Annu. Int. Conf. Digit. Government Res.*, May 2015, pp. 79–88.

[59] *Impact Assessment of Odine Programme*, Int. Data Corp. (IDC), Milan, Italy, 2019.

ROBERT ENRÍQUEZ-REYES received the master's degree from the Universidad Central del Ecuador, in 2012, the master's degree in management of communications and information systems from the Escuela Politécnica Nacional, in 2015, the Ph.D. degree in computer science from the University of Alicante, Spain, in 2019. He was a Specialist in oil and gas industry management at Saint Vincent College, USA, in 2013. He has experience in public and private national and international companies and is a Technology Manager in several companies of the finance and petroleum industry. He has taught for more than 22 years in several universities in Ecuador. He is currently a Professor with the Universidad Central del Ecuador (UCE). He has written more than 25 articles about technology management, security, and open data. He has directed 12 master's theses and has been the Director of the open data project at the Universidad Central del Ecuador. He is also a member of the UETIC and Cybersecurity projects sponsored by CEDIA. He is the General Workshop Chair of Industry 4.0 and Cybersecurity TIC EC CEDIA.

SUSANA CADENA-VELA received the Ph.D. degree in computer science in the line of data quality and open data. She is currently a Professor with the Universidad Central del Ecuador (UCE). She is also a member of the research groups, including Indicators for the Management of the Ecuadorian University, State of the IT of the Ecuadorian Universities sponsored by the Ecuadorian Consortium for the Development of Research and Academy (CEDIA), and the Group of Analytics and Big Data for the Cybersecurity, in addition to the groups Ecuadorian Network of Open Data and Metadata (REDAM) and the Open Science Research Group.

ANDRÉS FUSTER-GUILLÓ received the B.S. degree in computer science engineering from the Polytechnic University of Valencia, Spain, in 1995, and the Ph.D. degree in computer science from the University of Alicante, Spain, in 2003. Since 1997, he has been a member of the faculty of the Department of Computer Science and Technology, University of Alicante, where he is currently an Associate Professor. He was a Deputy Coordinator of the Polytechnic School and the Director of the Secretariat for Information Technology, University of Alicante. During this period, he has coordinated and participated in several strategic technological projects, including Open University (transparency portal and open data), UACloud, and Smart University, among others. He has published over 80 articles in different areas of research, including computer vision, 3D vision, machine learning, artificial neural networks, and open data.

JOSE-NORBERTO MAZÓN is currently an Associate Professor with the Department of Software and Computing Systems, University of Alicante, Spain. He is also the Chair of the Torrevieja's Venue of the University of Alicante. He is the author of more than 100 scientific publications in international conferences and journals. His research interests include open data, business intelligence in big data scenario, design of data-intensive web applications, smart cities, and smart tourism destinations.

LUIS DANIEL IBÁÑEZ is currently a Lecturer with the University of Southampton, U.K. He was a Deputy Coordinator of the QROWD and ODINE EU innovation actions from 2017 to 2019. His research is at the intersection between collaborative systems and data science, in particular, enabling collaboration on data acquisition, wrangling, and cleaning phases.

ELENA SIMPERL is currently a Professor of computer science with King's College London, UK. She is also a Fellow of the British Computer Society and a former Turing Fellow. From 2015 until early 2020, she led two European data incubators, ODINE and Data Pitch, helping almost a hundred small and medium businesses from more than 20 countries innovate with data. Her research is at the intersection between AI and crowd computing, helping designers understand how to build socio-technical systems that combine machine algorithms with human and social capabilities.